Six

Estimation

We now learn about a powerful use of statistics:

Statistical inference

about population parameters

using sample data.

In case you wonder about the relevance of learning about probability and sampling distribution, this is why:

- Statistical inference methods use probability calculations that assume that the data were gathered with a random sample or a randomized experiment.

- The probability calculations refer to a sampling distribution of a statistic, which is often approximately a normal distribution.

There are two types of statistical inference methods

- estimation of population parameters; and
- testing hypotheses about the parameter values.

This chapter discusses the first — estimating population parameters.

Two types of Estimations

Point estimation

Based on sample data, a single number is calculated to estimate the population parameter. The rule or formula that describes this calculation is called the point estimator. The resulting number is called a point estimate.

Interval estimation

Based on sample data, two numbers are calculated to form an interval within which the parameter is expected to lie.

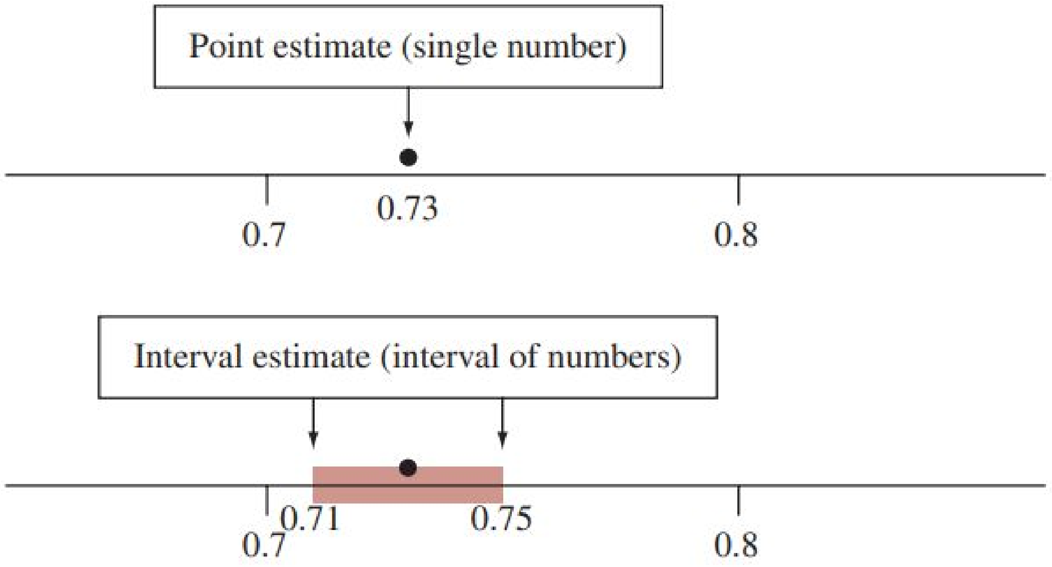
EXAMPLE 6.1

One survey asked, "Do you believe in hell?"

From sample data, the point estimate for the proportion of adult (in the population) who would respond "yes" is 0.73. The adjective "point" refers to using a single number as the parameter estimate.

An interval estimate predicts that the proportion of adult (in the population) who believe in hell falls between 0.71 and 0.75.

The next figure illustrates the difference between point estimate and interval estimate for the previous example.



1 Point Estimation

Suppose we are interested to estimate the parameter μ, the population mean. Assume that we have the following data, a random sample consisting

$$X_1, X_2, \ldots, X_n.$$

Definition 1 (Estimator)

*An **estimator** is a rule, usually expressed as a formula, that tells us how to calculate an **estimate** based on information in the sample.*

EXAMPLE 6.2 (POINT ESTIMATOR)

We want to estimate the average waiting time for a bus (μ) for students attending ST2334. The lecturer asked 4 students their waiting times $X_1, \ldots, X_4$ for a bus. The (observed) results are

$$x_1 = 6, x_2 = 1, x_3 = 4, x_4 = 9.$$

We can use $\overline{X} = \frac{1}{4}(X_1 + \cdots + X_4)$ to estimate μ. In this case, $\overline{X}$ is the estimator (for μ), and the computed value $\bar{x} = 5$ is the estimate.

QUESTIONS

- *How good is the estimator?*
- *What would be a criteria for a "good" estimator?*

Unbiased Estimator

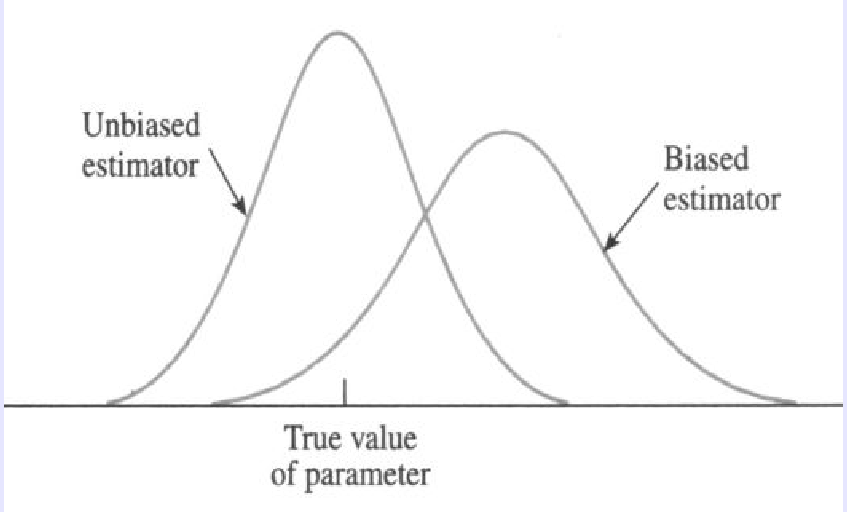
One of the reasons we think $\overline{X}$ is a good estimator of μ is because $E(\overline{X}) = \mu$. That is, “on average", the estimator is right.

In general, we represent the parameter of interest by θ. For example, θ can be p, μ, or σ.

DEFINITION 2 (UNBIASED ESTIMATOR)
*Let $\hat{\Theta}$ be an estimator of θ. Then $\hat{\Theta}$ is a random variable based on the sample. If $E(\hat{\Theta}) = \theta$, we call $\hat{\Theta}$ an **unbiased estimator** of θ.*

REMARK

An unbiased estimator has mean value equals to the true value of the parameter.



EXAMPLE 6.3 (UNBIASED ESTIMATOR)

Let $X_1, X_2, \ldots, X_n$ be a random sample from the same population with mean μ and variance σ^2. Then

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

is an unbiased estimator of σ^2 since $E(S^2) = \sigma^2$.

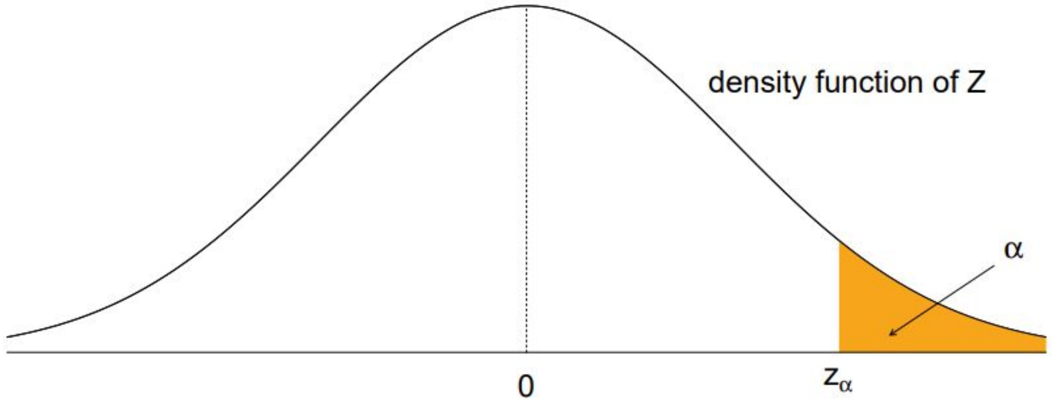
Maximum Error of Estimate

Typically $\overline{X} \neq \mu$, so $\overline{X} - \mu$ measures the difference between the estimator and the true value of the parameter.

Recall that if the population is normal or if n is large, $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ follows a standard normal or an approximately standard normal distribution.

DEFINITION 3 (z_α)

Define z_α to be the number with an upper-tail probability of α for the standard normal distribution Z. That is, $P(Z > z_\alpha) = \alpha$.

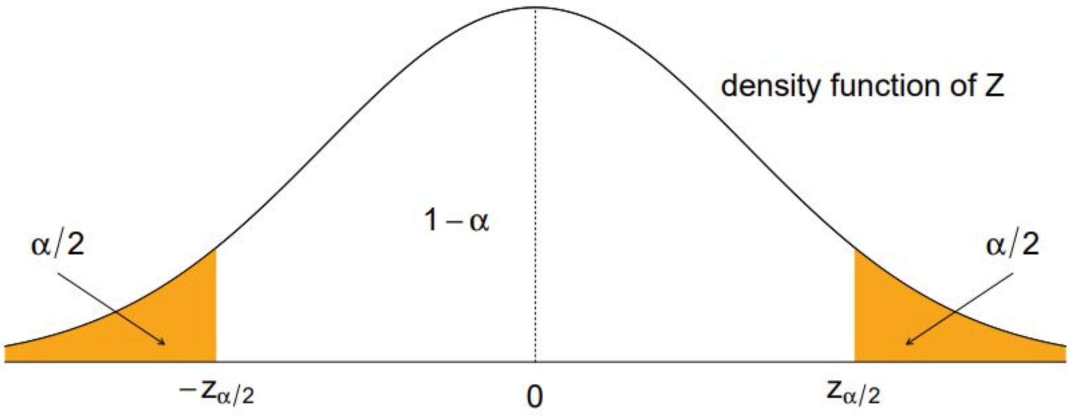


From the above definition, we then have

$$P\left(-z_{\alpha/2} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha.$$

In other words,

$$P\left(\frac{|\overline{X}-\mu|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = P\left(|\overline{X}-\mu| \leq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$



This means that, with probability $1-\alpha$, the error $|\overline{X}-\mu|$ is less than

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

Definition 4 (Maximum Error Of Estimate)

The quantity

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

is called the ***maximum error of estimate****.*

EXAMPLE 6.4 (TV TIME FOR INTERNET USERS)

An investigator is interested in the amount of time internet users spend watching television per week.

Based on historical experience, he assumes that the standard deviation is $\sigma = 3.5$ hours.

He proposes to select a random sample of $n = 50$ internet users, poll them, and take the sample mean to estimate the population mean μ.

What can he assert with probability 0.99 about the maximum error of estimate?

Solution:

As $n = 50 \geq 30$ is large, $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ is approximately normal.

So we can use the previous result, with $\sigma = 3.5$, $\alpha = 0.01$ and $z_{\alpha/2} = z_{0.005} = 2.576$.

With probability 0.99, the error is at most

$$E = 2.576 \times \frac{3.5}{\sqrt{50}} \approx 1.27.$$

Remark

$z_{0.005}$ is the same as the 0.995 quantile of the standard normal. The value of 2.576 can be obtained from tables or software.

Use the command `qnorm(0.995)` or `qnorm(0.005, lower.tail=F)` to obtain the value via https://rdrr.io/snippets/.

Alternatively, you may use Radiant to get the same value as well.

Determination of Sample Size

We often want to know what the minimum sample size should be, so that with probability $1-\alpha$, the error is at most E_0.

To answer this, consider the fact that we want

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq E_0.$$

Solving for n, we have

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{E_0}\right)^2.$$

Different Cases

We had previously understood the sampling distribution of $\overline{X}$ for a variety of cases. Repeating the same arguments above, we have the following table.

DIFFERENT CASES

	Population	σ	n	Statistic	E	n for desired E_0 and α
I	Normal	known	any	$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2} \cdot \sigma}{E_0}\right)^2$
II	any	known	large	$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2} \cdot \sigma}{E_0}\right)^2$
III	Normal	unknown	small	$T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$	$t_{n-1;\alpha/2} \cdot \frac{s}{\sqrt{n}}$	$\left(\frac{t_{n-1;\alpha/2} \cdot s}{E_0}\right)^2$
IV	any	unknown	large	$Z = \frac{\overline{X}-\mu}{S/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2} \cdot s}{E_0}\right)^2$

2 Confidence Intervals for the Mean

Since a point estimate is almost never right, one might be interested in asking for an interval where the parameter lies in.

Definition 5 (Confidence Interval)
*An **interval estimator** is a rule for calculating, from the sample, an interval (a,b) in which you are fairly certain the parameter of interest lies in.*

*This "fairly certain" can be quantified by the **degree of confidence** also known as **confidence level** $(1-\alpha)$, in the sense that*

$$P(a < \mu < b) = 1 - \alpha.$$

*(a,b) is called the $(1-\alpha)$ **confidence interval**.*

Case I: σ known, data normal

Consider the case where σ is known, and data comes from a normal population.

We learnt previously that

$$P\left(-z_{\alpha/2} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha.$$

Rearranging, we have

$$P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

So

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

is a $(1-\alpha)$ confidence interval.

EXAMPLE 6.5

In order to set inventory levels, a computer company samples demand during lead time over 25 time periods:

```
235 374 309 499 253 421 361 514 462 369 394 439
348 344 330 261 374 302 466 535 386 316 296 332 334
```

It is known that the (population) standard deviation of demand over lead time is 75 computers. Given that $\bar{x} = 370.16$, estimate the mean demand over lead time with 95% confidence. Assume a normal distribution for the population.

Solution:

Note that $z_{\alpha/2} = z_{0.025} = 1.96$. The 95% confidence interval is

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 370.16 \pm 1.96\frac{75}{\sqrt{25}} = 370.16 \pm 29.4$$

or $(340.76, 399.56)$.

Remark

Notice that our $(1-\alpha)$ confidence interval can be written as $\overline{X} \pm E$.

This is not a coincidence: recall that there is $(1-\alpha)$ confidence that the error $|\overline{X}-\mu|$ is within E.

For the other cases, based on our understanding of the sampling distribution of $\overline{X}$, we can construct our confidence intervals for the different cases $\overline{X} \pm E$, based on the conditions given.

Confidence Intervals for the Mean

The table below gives the $(1-\alpha)$ confidence interval (formulas) for the population mean.

Case	Population	σ	n	Confidence Interval
I	Normal	known	any	$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$
II	any	known	large	$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$
III	Normal	unknown	small	$\bar{x} \pm t_{n-1;\alpha/2} \cdot s/\sqrt{n}$
IV	any	unknown	large	$\bar{x} \pm z_{\alpha/2} \cdot s/\sqrt{n}$

Note that n is considered large when $n \geq 30$.

EXAMPLE 6.6 (WHICH CASE?)
The following data set collects $n = 41$ randomly sampled waiting times of students from ST2334 to receive reply for their email from a survey in the day time.

2.50	23.28	19.34	4.74	7.03	21.85	2.72
17.73	21.55	9.71	30.24	0.37	31.26	35.24
7.81	16.69	66.54	1.88	14.14	46.59	28.17
0.06	9.32	0.03	10.75	6.97	56.86	2.89
7.67	30.16	0.33	0.44	3.77	25.07	7.05
0.08	10.64	13.10	7.92	112.77	11.93	

Given that $\bar{x} = 17.736$ and $s = 21.7$, construct a 98% confidence interval for the mean waiting time of *all ST2334 students*.

Solution:

Note that σ is unknown, and n is large. So we are in Case IV.

Note that $z_{\alpha/2} = z_{0.01} = 2.326$. So our 98% confidence interval is

$$\begin{aligned} \bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} &= 17.736 \pm 2.326 \times \frac{21.7}{\sqrt{41}} \\ &= (9.85, 25.62). \end{aligned}$$

Example 6.7 (Which Case Again?)
The contents of 7 similar containers of sulphuric acid (in litres) are

```
9.8  10.2  10.4  9.8  10.0  10.2  9.6
```

It can shown that $\bar{x} = 10$ and $s^2 = 0.08$. Find a 95% confidence interval for the mean content of all such containers, assuming an approximate normal distribution for container contents.

Solution:
We are in Case III.

Using software, we obtain $t_{6;0.025} = 2.447$.

Thus a 95% confidence interval for the mean content of all such containers is given as

$$\bar{x} \pm t_{n-1;\alpha/2} \cdot \frac{s}{\sqrt{n}} = 10 \pm 2.447 \cdot \frac{\sqrt{0.08}}{\sqrt{7}} = (9.738, 10.262).$$

Interpreting Confidence Intervals I

- We saw that $\overline{X} \pm E$ has probability $(1-\alpha)$ of containing μ.

  This is a probability statement about the procedure by which we compute the interval — the interval estimator.

- Each time we take a sample, and go through this construction, we get a different confidence interval.

- Sometimes we get a confidence interval that contains μ, and sometimes we get one that does not contain μ.

- Once an interval is computed, μ is either in it or not. There is no more randomness.

Interpreting Confidence Intervals II

- Since μ is typically not known, there is no way to determine whether a particular confidence interval succeeded in capturing the population mean.
- However, if we repeat this procedure of taking a sample and computing a confidence interval many times, about $(1-\alpha)$ of the many confidence intervals that we get will contain the true parameter.

  This is what "confidence" means — a confidence in the method used.
- The following R Shiny app allows us to explore this fact:
  `https://istats.shinyapps.io/ExploreCoverage/`

3 Comparing Two Populations

In real applications, it is quite common to compare the means of two populations.

Imagine that we have two populations

- Population 1 has mean μ_1, variance σ_1^2.
- Population 2 has mean μ_2, variance σ_2^2.

Experimental Design

In order to compare two populations, a number of observations from each population need to be collected. Experimental design refers to the manner in which samples from populations are collected.

Two Basic Designs for comparing Two Treatments

- Independent samples — complete randomization.
- Matched pairs samples — randomization between matched pairs.

Example 6.8 (Independent Samples)
In order to compare the examination scores of male and female students attending ST2334,

- 10 scores of female students are randomly sampled — Sample I,
- 8 scores of male students are randomly sampled — Sample II.

Note that all observations are independent —

- Sample I and Sample II are independent;
- Individuals within Sample I are independent;
- Individuals within Sample II are independent.

EXAMPLE 6.9 (MATCHED PAIRS SAMPLES)

In order to study whether there exists income difference between male and female, 100 married couples are sampled, and their monthly incomes are collected.

In this example, the treatment groups are the female group and male group.

Note that observations are dependent in a special way —

- Within the pair, the observations are dependent (since they are married to one another);
- Between pairs, observations are independent.

4 Independent Samples: Unequal variances

Our interest is to make statistical inference on $\mu_1 - \mu_2$. Consider the following assumptions:

Independent Samples (Known and unequal variances)

1. A random sample of size n_1 from population 1 with mean μ_1 and variance σ_1^2.
2. A random sample of size n_2 from population 2 with mean μ_2 and variance σ_2^2.
3. The two samples are independent.
4. The population variances are known and not the same: $\sigma_1^2 \neq \sigma_2^2$
5. Either one of the following conditions holds:
    - The two populations are normal; OR
    - Both samples are large: $n_1 \geq 30$, $n_2 \geq 30$.

Consider $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$, random samples from the two populations of interest. Let

$$\overline{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i, \text{ and } \overline{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i$$

be the means of random samples. Then,

$$E(\overline{X}) = \mu_1, \quad V(\overline{X}) = \frac{\sigma_1^2}{n_1}, \quad E(\overline{Y}) = \mu_2, \quad V(\overline{Y}) = \frac{\sigma_2^2}{n_2}.$$

Thus

$$E(\overline{X} - \overline{Y}) = \mu_1 - \mu_2 = \delta,$$

and, using the independence assumption,

$$V(\overline{X} - \overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

When

- the two populations are normal, OR
- both samples are large,

we have

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0,1).$$

Confidence Intervals for $\mu_1 - \mu_2$

We are interested in the difference

$$\delta = \mu_1 - \mu_2,$$

with confidence $100(1-\alpha)\%$ for any $0 < \alpha < 1$.

If σ_1^2 and σ_2^2 are **known**, by the distributions above, we have

$$P\left(\left|\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right| < z_{\alpha/2}\right) = 1-\alpha$$

or

$$P\left((\overline{X}-\overline{Y})-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} < \mu_1-\mu_2 < (\overline{X}-\overline{Y})+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right) = 1-\alpha.$$

Thus the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\left((\overline{X}-\overline{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\ (\overline{X}-\overline{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \right)$$

or

$$(\overline{X}-\overline{Y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}.$$

CONFIDENCE INTERVALS: KNOWN AND UNEQUAL VARIANCES

Suppose we have independent populations with known and unequal variances, and that either one of the following conditions holds:

- The two populations are normal; OR
- Both samples are large: $n_1 \geq 30$, $n_2 \geq 30$.

The $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$, is then given as

$$(\bar{x} - \bar{y}) \pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}.$$

EXAMPLE 6.10

A study was conducted to compare two types of engines, A and B.

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.

- 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 the 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:
For a 96% confidence interval, $\alpha = 0.04$ and $z_{0.02} = 2.05$. We are also given that

$$n_1 = 50, \bar{x}_A = 36,\ \sigma_1^2 = 6^2$$
$$n_2 = 75, \bar{x}_B = 42,\ \sigma_2^2 = 8^2$$

The sample sizes are large, so a 96% confidence interval for $\mu_B - \mu_A$ is

$$\begin{aligned}
&(\bar{x}_B - \bar{x}_A) \pm z_{\alpha/2}\sqrt{\sigma_2^2/n_2 + \sigma_1^2/n_1} \\
&= (42 - 36) \pm 2.05 \cdot \sqrt{8^2/75 + 6^2/50} \\
&= (3.428, 8.571).
\end{aligned}$$

We next consider the following assumptions/case:

INDEPENDENT SAMPLES (LARGE, WITH UNKNOWN VARIANCES)

1. A random sample of size n_1 from population 1 with mean μ_1 and variance σ_1^2.
2. A random sample of size n_2 from population 2 with mean μ_2 and variance σ_2^2.
3. The two samples are independent.
4. The population variances are unknown and not the same: $\sigma_1^2 \neq \sigma_2^2$
5. Both samples are large: $n_1 \geq 30$, $n_2 \geq 30$.

Since σ_1 and σ_2 are unknown, let

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2, \text{ and } S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$$

and use

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx N(0, 1).$$

If σ_1^2 and σ_2^2 are **unknown**, the $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is

$$\left((\overline{X} - \overline{Y}) - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},\ (\overline{X} - \overline{Y}) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right)$$

or

$$(\overline{X} - \overline{Y}) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

Confidence Intervals: Large, with unknown variances

Suppose we have independent populations with unknown and unequal variances, and that both samples are large: $n_1 \geq 30$, $n_2 \geq 30$.

The $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$, is then given as

$$(\overline{x} - \overline{y}) \pm z_{\alpha/2}\sqrt{s_1^2/n_1 + s_2^2/n_2}.$$

5 Independent Samples: Equal variances

Consider the following assumptions:

INDEPENDENT SAMPLES: SMALL, WITH EQUAL VARIANCES

1. A random sample of size n_1 from population 1 with mean μ_1 and variance σ_1^2.
2. A random sample of size n_2 from population 2 with mean μ_2 and variance σ_2^2.
3. The two samples are independent.
4. The population variances are unknown and the same: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
5. Both samples are small: $n_1 < 30$, $n_2 < 30$
6. Both populations are normally distributed.

The Equal Variance Assumption

In real applications, the equal variance assumption is usually unknown and needs to be checked.

Based upon the normal distribution and equal variance assumptions

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1).$$

Since σ is unknown, we shall estimate it.

Note that S_1^2 and S_2^2 are both unbiased estimators of σ^2 under the equal variance assumption.

We can use the pooled estimator to estimate σ^2 better.

Definition 6 (The Pooled estimator: S_p^2)
*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.

When we estimate σ^2 using S_p^2, the resulting statistic

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

follows a t-distribution with degrees of freedom $n_1 + n_2 - 2$.

We then have

$$P\left(-t_{n_1+n_2-2;\alpha/2} < \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < t_{n_1+n_2-2;\alpha/2}\right) = 1-\alpha.$$

Confidence Intervals: Small, with equal variances

Suppose we have independent, normal populations with unknown and equal variances, and that both samples are small: $n_1 < 30$, $n_2 < 30$.

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

$$(\overline{X} - \overline{Y}) \pm t_{n_1+n_2-2;\alpha/2}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

EXAMPLE 6.11

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.
- 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:
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 $\bar{x} - \bar{y} = 85 - 81 = 4$ is the point estimate for $\mu_1 - \mu_2$.

As we assume equal population variance, we estimate it by the pooled variance

$$s_p^2 = \frac{(12-1) \times 4^2 + (10-1) \times 5^2}{12+10-2} = 20.05.$$

In this case, $t_{n_1+n_2-2;\alpha/2} = t_{20;0.05} = 1.7247$. Thus a 90% confidence interval for $\mu_1 - \mu_2$ is given as

$$
\begin{aligned}
&(\overline{x} - \overline{y}) \pm t_{n_1+n_2-2;\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\
&= (85 - 81) \pm 1.7247 \times \sqrt{20.05} \times \sqrt{\frac{1}{12} + \frac{1}{10}} \\
&= (0.693, 7.307).
\end{aligned}
$$

Independent Large Samples with Equal Variance

Note that for large samples such that $n_1 \geq 30$, $n_2 \geq 30$, we can replace $t_{n_1+n_2-2;\alpha/2}$ by $z_{\alpha/2}$ in the previous formula.

Confidence Intervals: Large, with equal variances

Suppose we have independent populations with unknown and equal variances, and that both samples are large: $n_1 \geq 30$, $n_2 \geq 30$.

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

$$(\overline{X} - \overline{Y}) \pm z_{\alpha/2}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

6 PAIRED DATA

Some times, like in the couple income example, it makes sense to take matched pairs instead of independent samples.
Because of dependence in the sample, the methods discussed previously are not applicable.

Consider the assumptions that follows.

Paired Data

1. $(X_1, Y_1), \ldots, (X_n, Y_n)$ are matched pairs, where $X_1, \ldots, X_n$ is a random sample from population 1, $Y_1, \ldots, Y_n$ is a random sample from population 2.
2. X_i and Y_i are dependent.
3. (X_i, Y_i) and (X_j, Y_j) are independent for any $i \neq j$.
4. For matched pairs, define $D_i = X_i - Y_i, \quad \mu_D = \mu_1 - \mu_2$.
5. Now we can treat $D_1, D_2, \ldots, D_n$ as a random sample from a single population with mean μ_D and variance σ_D^2.

All techniques derived for a single population can now be employed.

- We consider the statistic

$$T = \frac{\overline{D}-\mu_D}{S_D/\sqrt{n}}, \quad \text{where} \quad \overline{D} = \frac{\sum_{i=1}^n D_i}{n}, \quad S_D^2 = \frac{\sum_{i=1}^n (D_i-\overline{D})^2}{n-1}.$$

- If $n < 30$ and the population is normally distributed then

$$T \sim t_{n-1}.$$

- If $n \geq 30$, then

$$T \sim N(0,1).$$

Confidence Intervals: Paired Data

For paired data, if n is small ($n < 30$) and the population is normally distributed, a $(1-\alpha)100\%$ confidence interval for μ_D is

$$\overline{d} \pm t_{n-1;\alpha/2} \cdot \frac{s_D}{\sqrt{n}}.$$

If n is large ($n \geq 30$), a $(1-\alpha)100\%$ confidence interval for μ_D is

$$\overline{d} \pm z_{\alpha/2} \cdot \frac{s_D}{\sqrt{n}}.$$

Example 6.12

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.

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

Given that $\overline{d} = -1.6$ and $s_D^2 = 40.71$, compute a 98% confidence interval for the true difference in the two learning procedures.

Solution:
Since $\alpha = 0.02$, we have $t_{n-1;\alpha/2} = t_{9,0.01} = 2.821$. Thus a 98% confidence interval for the true difference μ_D is given as

$$\overline{d} \pm t_{n-1;\alpha/2} \cdot \frac{s_D}{\sqrt{n}} = -1.6 \pm 2.821 \times \sqrt{\frac{40.71}{10}} = (-7.292, 4.092).$$