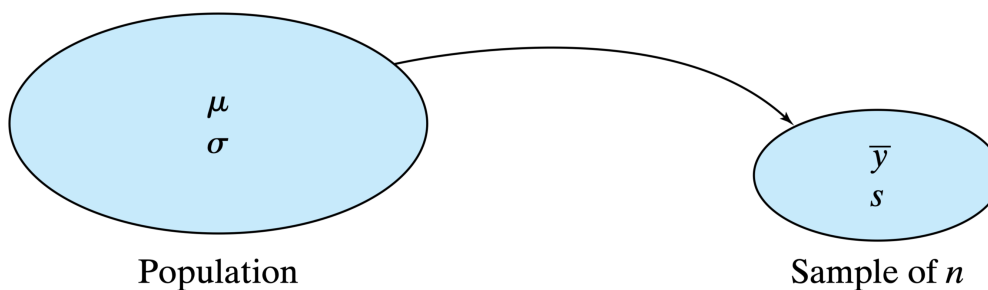


# Confidence Intervals

In this chapter we undertake our first substantial adventure into statistical inference. Recall that statistical inference is based on the **random sampling model**: *We view our data as a random sample from some population, and we use the information in the sample to infer facts about the population.*

Statistical estimation is a form of statistical inference. We will learn how to assess the precision of the estimate.

In general, for a sample of observations on a quantitative variable  $Y$ , the sample mean and SD are estimates of the population mean and SD:



Our goal is to estimate  $\mu$ . We will see how to assess the reliability or precision of this estimate, and how to plan a study large enough to attain a desired precision.

## Example: butterfly wings

As part of a larger study of body composition, researchers captured 14 male Monarch butterflies at Oceano Dunes State Park in California and measured wing area (in  $\text{cm}^2$ ). The data are given in the following table

Table 6.1.1 Wing areas of male monarch butterflies				
Wing area ( $\text{cm}^2$ )				
33.9	33.0	30.6	36.6	36.5
34.0	36.1	32.0	28.0	32.0
32.2	32.2	32.3	30.0	

For these data, the mean and standard deviation are  $\bar{y} = 32.81 \text{ cm}^2$  and  $s = 2.48 \text{ cm}^2$ . Define the population mean and SD as follows:

- $\mu$  = the (population) mean wing area of male Monarch butterflies in the Oceano Dunes region;
- $\sigma$  = the (population) SD of wing area of male Monarch butterflies in the Oceano Dunes region.

It is natural to estimate  $\mu$  by the sample mean and  $\sigma$  by the sample SD. Specifically,

- 32.81 is an estimate of  $\mu$ ;
- 2.48 is an estimate of  $\sigma$ .

These estimates are subject to sampling error (not only measurement error). The task of this chapter is to assess the reliability or precision of  $\bar{y}$ .

## Standard Error of the Mean

The standard deviation of the sampling distribution of  $\bar{Y}$  is

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}.$$

The population standard deviation  $\sigma$  is typically unknown. Since  $s$  is an estimate of  $\sigma$ , a natural estimate of  $\sigma/\sqrt{n}$  would be

$$\text{SE} = \text{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}},$$

which is called the standard error of the mean.

For the butterfly wings example, the standard error of the mean is

$$\text{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{2.48}{\sqrt{14}} = 0.66 \text{ cm}^2.$$

## Standard error $s/\sqrt{n}$ versus sample standard deviation $s$

The sample SD  $s$  describes the dispersion of the data, while the SE

$$\frac{s}{\sqrt{n}}$$

describes the unreliability (due to sampling error) in the mean of the sample as an estimate of the mean of the population.

## Example: lamb birthweights

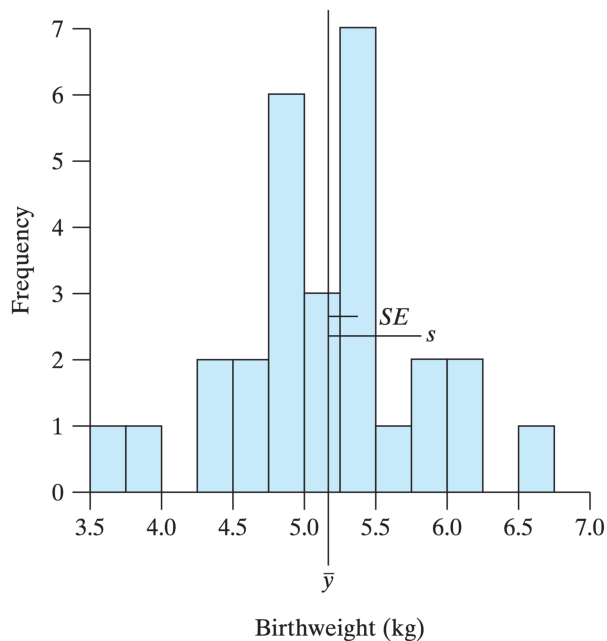
A geneticist weighed 28 female lambs at birth. The lambs were all born in April, were all the same breed (Rambouillet), and were all single births (no twins). The diet and other

environmental conditions were the same for all the parents. The birthweights are shown in the following table.

Table 6.2.1 Birthweights of 28 Rambouillet lambs						
Birthweight (kg)						
4.3	5.2	6.2	6.7	5.3	4.9	4.7
5.5	5.3	4.0	4.9	5.2	4.9	5.3
5.4	5.5	3.6	5.8	5.6	5.0	5.2
5.8	6.1	4.9	4.5	4.8	5.4	4.7

For these data, the mean is  $\bar{y} = 5.17$  kg, the sample SD is  $s = 0.65$  kg, and the SE is  $SE = 0.12$  kg. The sample SD,  $s$ , describes *the variability of birthweights among the lambs in the sample*, while the SE indicates *the variability associated with the sample mean (5.17 kg)*, viewed as an estimate of the population mean birthweight.

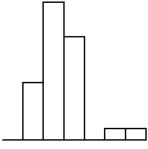
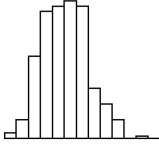
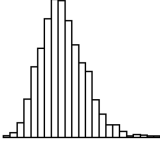
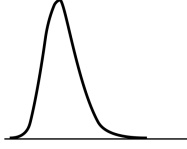
This distinction is emphasized in the figure below, which shows a histogram of the lamb birthweight data; the sample SD is indicated as a deviation from the sample mean  $\bar{y}$ , while the SE is indicated as variability associated with  $\bar{y}$  itself.



- For very large  $n$ , the sample mean and SD  $\bar{y}$  and  $s$  would be very close to the population mean and SD  $\mu$  and  $\sigma$ .
- The SE, by contrast, tends to decrease as  $n$  increases; when  $n$  is very large, the SE is very small and so the sample mean is a very precise estimate of the population mean.

	$n = 28$	$n = 280$	$n = 2,800$	$n \rightarrow \infty$
$\bar{y}$	5.17	5.19	5.14	$\bar{y} \rightarrow \mu$
$s$	0.65	0.67	0.65	$s \rightarrow \sigma$
SE	0.12	0.040	0.012	SE $\rightarrow 0$

Sample distribution	$n = 28$	$n = 280$	$n = 2,800$	$n \rightarrow \infty$
				

```
In [56]: # generate random sample of various sizes from
# standard normal distribution
res <- matrix(nrow = 4, ncol = 3)
colnames(res) <- c('sample mean', 'sample SD', 'SE')
y <- rnorm(10)
res[1, ] <- round(c(mean(y), sd(y), sd(y)/sqrt(10)), 2)
y <- rnorm(100)
res[2, ] <- round(c(mean(y), sd(y), sd(y)/sqrt(100)), 2)
y <- rnorm(1000)
res[3, ] <- round(c(mean(y), sd(y), sd(y)/sqrt(1000)), 2)
y <- rnorm(10000)
res[4, ] <- round(c(mean(y), sd(y), sd(y)/sqrt(10000)), 2)
res
```

A matrix: 4 × 3 of type dbl

sample mean	sample SD	SE
-0.19	1.20	0.38
-0.04	1.09	0.11
-0.03	1.00	0.03
0.01	1.00	0.01

## Graphical representation of the SE and the sample SD

```
In [52]: library(ggplot2)

f <- function(x) {
  c(mean(x), sd(x), sd(x)/sqrt(length(x)))
}
y <- aggregate(iris$Sepal.Length, list(iris$Species), FUN = f)
df <- data.frame(species = y[, 1])
df[, 2:4] <- as.data.frame(rbind(y[1, 2], y[2, 2], y[3, 2]))
names(df)[2:4] <- c('mean', 'sd', 'se')

g1 <- ggplot(df, aes(x = species, y = mean)) +
  geom_point(stat="identity", fill = 'skyblue') +
  geom_errorbar(aes(ymin = mean - sd, ymax = mean + sd), width = .2) +
  labs(title = "Interval plots of Sepal Length by Species",
       x = "Species", y = "Sepal Length") +
  scale_y_continuous(limits = c(0, 8)) +
  theme_bw() +
  theme(text = element_text(size = 15))
g2 <- ggplot(df, aes(x = species, y = mean)) +
  geom_bar(stat="identity", fill = 'skyblue') +
```

```

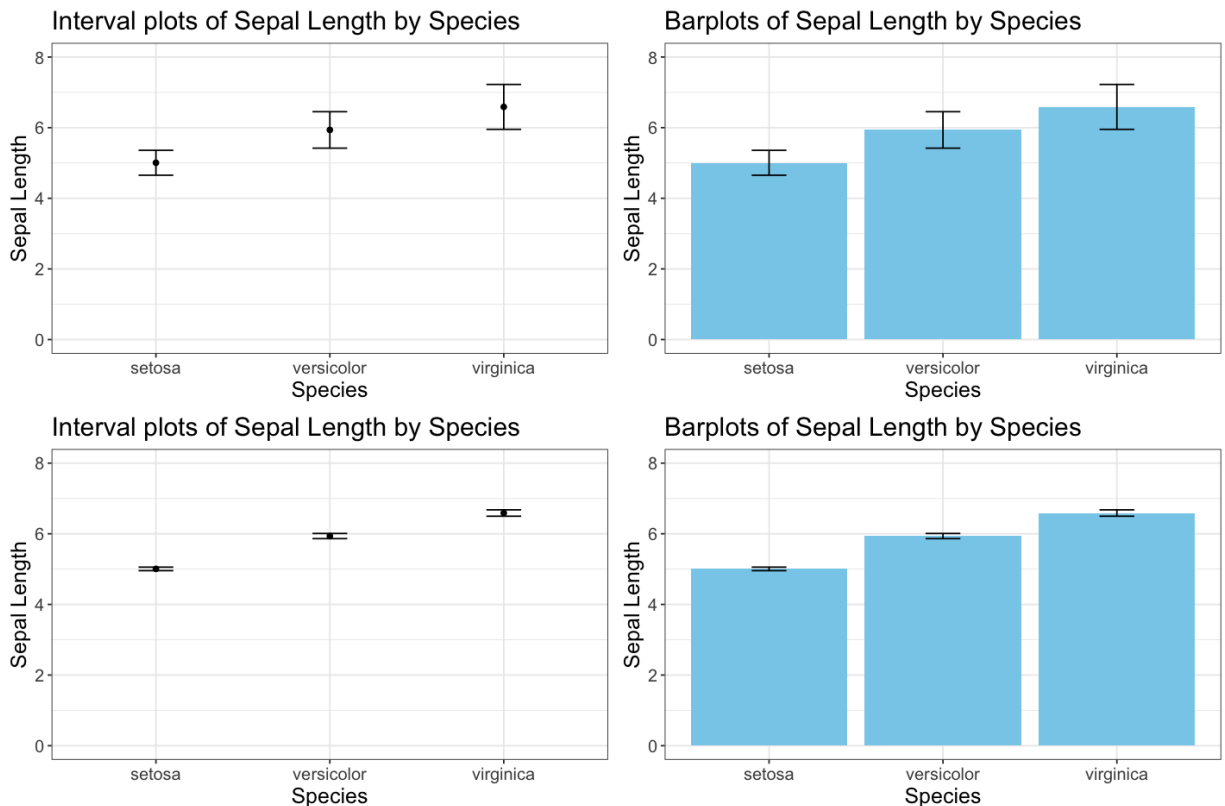
geom_errorbar(aes(ymin = mean - sd, ymax = mean + sd), width = .2) +
labs(title = "Barplots of Sepal Length by Species",
      x = "Species", y = "Sepal Length") +
scale_y_continuous(limits = c(0, 8)) +
theme_bw() +
theme(text = element_text(size = 15))
g3 <- ggplot(df, aes(x = species, y = mean)) +
geom_point(stat="identity", fill = 'skyblue') +
geom_errorbar(aes(ymin = mean - se, ymax = mean + se), width = .2) +
labs(title = "Interval plots of Sepal Length by Species",
      x = "Species", y = "Sepal Length") +
scale_y_continuous(limits = c(0, 8)) +
theme_bw() +
theme(text = element_text(size = 15))
g4 <- ggplot(df, aes(x = species, y = mean)) +
geom_bar(stat="identity", fill = 'skyblue') +
geom_errorbar(aes(ymin = mean - se, ymax = mean + se), width = .2) +
labs(title = "Barplots of Sepal Length by Species",
      x = "Species", y = "Sepal Length") +
scale_y_continuous(limits = c(0, 8)) +
theme_bw() +
theme(text = element_text(size = 15))

```

```

In [53]: library(patchwork)
options(repr.plot.width=12, repr.plot.height=8)
(g1 + g2)/(g3 + g4)

```



## Confidence Interval for $\mu$

The standard error of the mean (the SE) measures how far  $\bar{y}$  is likely to be from the population mean  $\mu$ . In this section we make this idea precise.

- Parameter  $\mu$  denotes the population mean, which is **fixed** but unknown.
- The sample mean  $\bar{Y}$  (**random**) is an estimate of  $\mu$ , whose performance can be evaluated by its sampling distribution.
- A confidence interval for the population mean is a range of values within which we expect the true population mean to fall with a certain level of confidence.
- Recall that for  $Z \sim N(0, 1)$ , the probability that  $Z$  is between  $\pm 2$  is about 95%. More precisely,  $P(-1.96 < Z < 1.96) = 0.95$ .
- From Chapter 5 we know that if the population  $Y$  has a normal distribution, then the sampling distribution of  $\bar{Y}$  is  $N(\mu, \sigma^2/n)$ , so

$$P(-1.96 < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < 1.96) = 0.95.$$

- Simplifying the above equation leads to

$$P(\bar{Y} - 1.96 \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + 1.96 \times \frac{\sigma}{\sqrt{n}}) = 0.95.$$

- That is, the interval

$$\bar{Y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

will contain the population mean  $\mu$  for 95% of all samples.

- The interval

$$\bar{Y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

cannot be used for data analysis since the population SD  $\sigma$  is typically unknown.

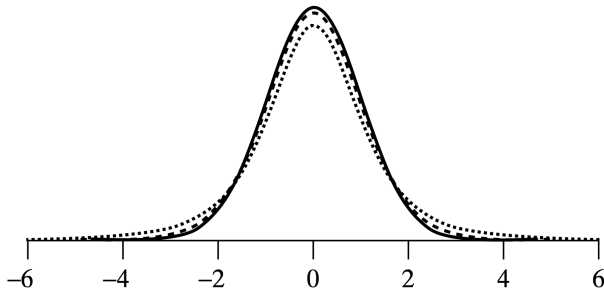
- If we replace  $\sigma$  by its estimate  $s$ , then we can calculate an interval from the data, but what happens to the 95% interpretation?
- It turns out that

$$\frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t_{n-1},$$

where  $t_{n-1}$  denotes the Student's  $t$  distribution with degrees of freedom  $\text{df} = n - 1$ .

- A  $t$  curve is **symmetric and bell shaped** like the normal curve but has a larger standard deviation (**heavier tail**). As the  $\text{df}$  increases, the  $t$  curves approach the normal curve; thus, the normal curve can be regarded as a  $t$  curve with infinite  $\text{df}$  ( $\text{df} = \infty$ ).

- The quantity  $(n - 1)$  is referred to as "degrees of freedom" because the deviations  $(Y_i - \bar{Y})$  must sum to zero, and so only  $(n - 1)$  of them are "free" to vary. A sample of size  $n$  provides only  $(n - 1)$  independent pieces of information about variability, that is, about  $\sigma$ .

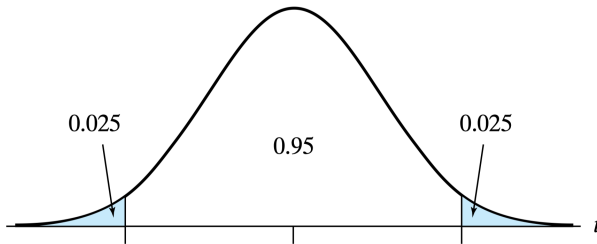


By the symmetry of the Student's  $t$  distribution and the fact that

$$\frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t_{n-1},$$

one has

$$P(-t_{n-1}(0.025) < \frac{\bar{Y} - \mu}{s/\sqrt{n}} < t_{n-1}(0.025)) = 0.95.$$



- The quantity  $t_{n-1}(0.025)$  is called the "two-sided 5% critical value" of Student's  $t$  distribution.
- Critical values of Student's  $t$  distribution are tabulated in  $t$  Table (Files --> Tables).
- The values of  $t_{n-1}(0.025)$  are shown in the row headed "df  $n - 1$ " and the column headed "Upper Tail Probability 0.025."
- $t$  Table v.s. Z Table.

Simplifying the following equation

$$P(-t_{n-1}(0.025) < \frac{\bar{Y} - \mu}{s/\sqrt{n}} < t_{n-1}(0.025)) = 0.95$$

leads to

$$P(\bar{Y} - t_{n-1}(0.025) \times \frac{s}{\sqrt{n}} < \mu < \bar{Y} + t_{n-1}(0.025) \times \frac{s}{\sqrt{n}}) = 0.95.$$

The interval

$$\bar{Y} \pm t_{n-1}(0.025) \times \frac{s}{\sqrt{n}}$$

is called the (two-sided) 95% confidence interval (CI) for  $\mu$ .

Generally, the two-sided  $1 - \alpha$  confidence interval for  $\mu$  is constructed using  $t_{n-1}(\alpha/2)$  as follows:

$$\bar{Y} \pm t_{n-1}(\alpha/2) \times \frac{s}{\sqrt{n}}.$$

### Example: butterfly wings

For the butterfly data, we have  $n = 14$ ,  $\bar{Y} = 32.8143 \text{ cm}^2$ , and  $s = 2.4757 \text{ cm}^2$ . Find a two-sided 95% confidence interval for the population mean  $\mu$ .

- $\alpha = 1 - 95\% = 0.05$ .
- According to the formula, the two-sided 95% confidence interval for  $\mu$  is

$$32.8143 \pm t_{14-1}(0.05/2) \times \frac{2.4757}{\sqrt{14}}.$$

- From  $t$  Table we find  $t_{13}(0.025) = 2.160$ .
- It follows that the two-sided 95% confidence interval for  $\mu$  is  $32.81 \pm 1.43$ ; that is (31.4, 34.2).
- The confidence statement asserts that the population mean wing area of male Monarch butterflies in the Oceano Dunes region of California is between  $31.4 \text{ cm}^2$  and  $34.2 \text{ cm}^2$  with 95% confidence.

Find a two-sided 90% confidence interval for the population mean  $\mu$ .

- $\alpha = 1 - 90\% = 0.1$ .
- According to the formula, the two-sided 90% confidence interval for  $\mu$  is

$$32.8143 \pm t_{14-1}(0.1/2) \times \frac{2.4757}{\sqrt{14}}.$$

- From  $t$  Table we find  $t_{13}(0.05) = 1.771$ .
- It follows that the two-sided 90% confidence interval for  $\mu$  is  $32.81 \pm 1.1718$ ; that is (31.6, 34.0).
- The confidence statement asserts that the population mean wing area of male Monarch butterflies in the Oceano Dunes region of California is between  $31.6 \text{ cm}^2$  and  $34.0 \text{ cm}^2$  with 90% confidence.
- *The higher the confidence level, the wider the confidence interval (for a fixed sample size; but note that as  $n$  increases the intervals tend to get smaller).*

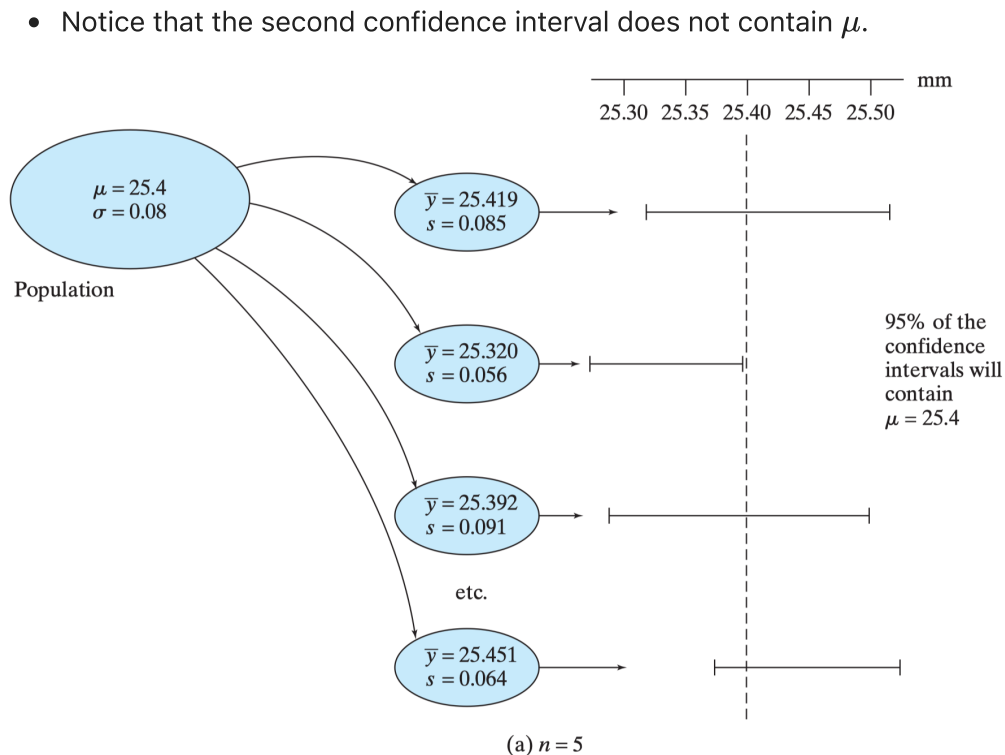
### Interpretation of a confidence interval



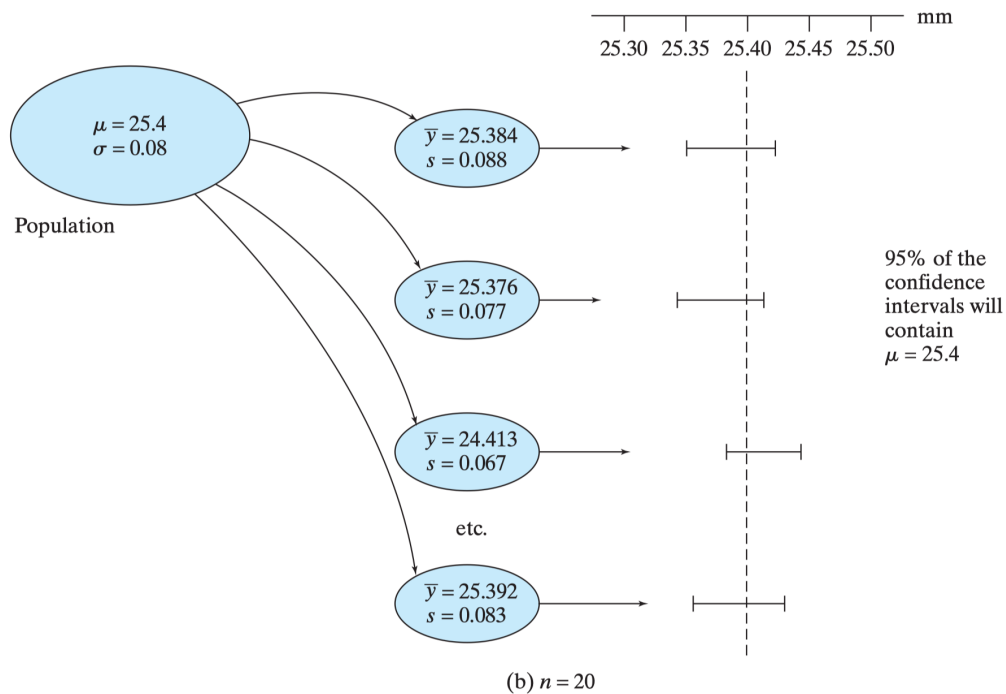
- Consider, for instance, a 95% confidence interval. One way to interpret the confidence level (95%) is to refer to the **meta-study** of repeated samples from the same population.
- If a 95% confidence interval for  $\mu$  is constructed for each sample, then 95% of the confidence intervals will contain  $\mu$ .
- Of course, the observed data in an experiment comprise only one of the possible samples; we can hope "confidently" that this sample is one of the lucky 95%, but we will never know.

## Example: blue jay bill length

In a certain large population of Blue Jays, the distribution of bill lengths is normal with mean  $\mu = 25.4$  mm and standard deviation  $\sigma = 0.08$  mm. Figure below shows some typical samples from this population; plotted on the right are the associated 95% confidence intervals. The sample sizes are  $n = 5$  and  $n = 20$ .



- In the totality of potential confidence intervals, the percentage that would contain  $\mu$  is 95% for either sample size.
- The larger samples tend to produce narrower confidence intervals.



A confidence level can be interpreted as a probability, but caution is required. If we consider 95% confidence intervals, for instance, then the following statement is correct:

$$P(\text{the next sample will give us a confidence interval that contains } \mu) = 0.95.$$

However, one should realize that *it is the confidence interval that is the random item* in this statement, and *it is not correct to replace this item with its value from the data*. Thus, for instance, we found in the butterfly wings example that the 95% confidence interval for the mean butterfly wings is

$$31.4 \text{ cm}^2 < \mu < 34.2 \text{ cm}^2.$$

Nevertheless, it is **not** correct to say that

$$P(31.4 \text{ cm}^2 < \mu < 34.2 \text{ cm}^2) = 0.95.$$

because this statement has no chance element; *either  $\mu$  is between 31.4 and 34.2 or it is not*.

## Example: bone mineral density

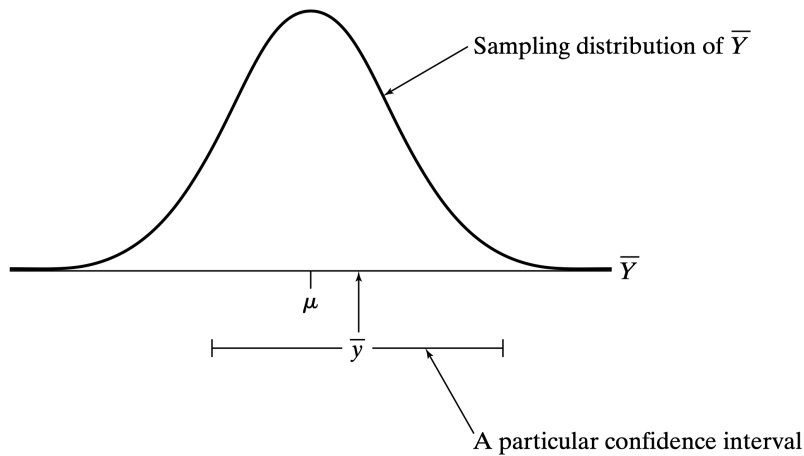
In an experiment to assess the effectiveness of hormone replacement therapy, researchers gave conjugated equine estrogen (CEE) to a sample of 94 women between the ages of 45 and 64. After taking the medication for 36 months, the bone mineral density was measured for each of the 94 women. The average density was 0.878 g/cm<sup>2</sup>, with a standard deviation of 0.126 g/cm<sup>2</sup>. Assume the bone mineral density is normally distributed.

- A two-sided 95% confidence interval for  $\mu$  is

$$0.878 \pm t_{94-1}(0.05/2) \times \frac{0.126}{\sqrt{94}}.$$

- According to  $t$  Table,  $t_{93}(0.025) = 1.984$  (since  $t$  Table doesn't list 93 degrees of freedom, we use 100 degrees of freedom).
- It follows that the two-sided 95% confidence interval for  $\mu$  is (0.852, 0.904).
- Thus, we are 95% confident that the mean bone mineral density of all women age 45 to 64 who take CEE for 36 months is between 0.852 g/cm<sup>2</sup> and 0.904 g/cm<sup>2</sup>.

## Relationship to sampling distribution of $\bar{Y}$



Notice that the particular confidence interval does contain  $\mu$ ; this will happen for 95% of samples.

## One-sided confidence intervals

- Most confidence intervals are of the form "estimate  $\pm$  margin of error"; these are known as two-sided intervals.
- However, it is possible to construct a one-sided confidence interval, which is appropriate when only a lower bound, or only an upper bound, is of interest.

### Upper one-sided confidence intervals

- Simplifying the following equation

$$P(-t_{n-1}(\alpha) < \frac{\bar{Y} - \mu}{s/\sqrt{n}} < \infty) = 1 - \alpha$$

leads to

$$P(-\infty < \mu < \bar{Y} + t_{n-1}(\alpha) \times \frac{s}{\sqrt{n}}) = 1 - \alpha.$$

The interval

$$\left(-\infty, \bar{Y} + t_{n-1}(\alpha) \times \frac{s}{\sqrt{n}}\right)$$

is called the upper one-sided  $1 - \alpha$  confidence interval for  $\mu$ .

### Lower one-sided confidence intervals

- Simplifying the following equation

$$P\left(-\infty < \frac{\bar{Y} - \mu}{s/\sqrt{n}} < t_{n-1}(\alpha)\right) = 1 - \alpha$$

leads to

$$P\left(\bar{Y} - t_{n-1}(\alpha) \times \frac{s}{\sqrt{n}} < \mu < \infty\right) = 1 - \alpha.$$

The interval

$$\left(\bar{Y} - t_{n-1}(\alpha) \times \frac{s}{\sqrt{n}}, \infty\right)$$

is called the lower one-sided  $1 - \alpha$  confidence interval for  $\mu$ .

### Example: seeds per fruit

The number of seeds per fruit for the freshwater plant *Vallisneria americana* varies considerably from one fruit to another. A researcher took a random sample of 12 fruit and found that the average number of seeds was 320, with a standard deviation of 125. The researcher expected the number of seeds to follow, at least approximately, a normal distribution.

It might be that we want a lower bound on  $\mu$ , the population mean, but we are not concerned with how large  $\mu$  might be --> Lower one-sided confidence intervals

- For  $\alpha = 0.05$ , the lower limit of the confidence interval is

$$\bar{Y} - t_{n-1}(\alpha) \times \frac{s}{\sqrt{n}} = 320 - 1.796 \times 36 = 255.$$

The lower one-sided 95% confidence interval is thus  $(255, \infty)$  and we are 95% confident that the (population) mean number of seeds per fruit for *Vallisneria americana* is at least 255.

### Planning a Study to Estimate $\mu$

- Recall that as an estimate of the population mean  $\mu$ ,  $\bar{Y}$  has the sampling distribution with mean  $\mu$  and SD  $\sigma/\sqrt{n}$ . The precision with which a population mean  $\mu$  can be

estimated is thus determined by two factors: (1) *the population variability of the observed variable  $Y$* , and (2) *the sample size*.

- Suppose that plans have been made to reduce the variability of  $Y$  as much as possible, or desirable. What sample size will be sufficient to achieve a desired degree of precision in estimation of the population mean?
- If we use the standard error as our measure of precision, then this question becomes: What sample size will be sufficient to make the following inequality hold?

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}} \leq \text{Desired SE.}$$

## Example: butterfly wings

The butterfly wing data yielded the following summary statistics:

$$\bar{Y} = 32.81 \text{ cm}^2, s = 2.48 \text{ cm}^2, SE = 0.66 \text{ cm}^2$$

Suppose the researcher is now planning a new study of butterflies and has decided that it would be desirable that the SE be no more than  $0.4 \text{ cm}^2$ . As a preliminary guess of the SD, she will use the value from the old study, namely  $2.48 \text{ cm}^2$ . Thus, the desired sample size  $n$  must satisfy the following relation:

$$SE = \frac{2.48}{\sqrt{n}} \leq 0.4.$$

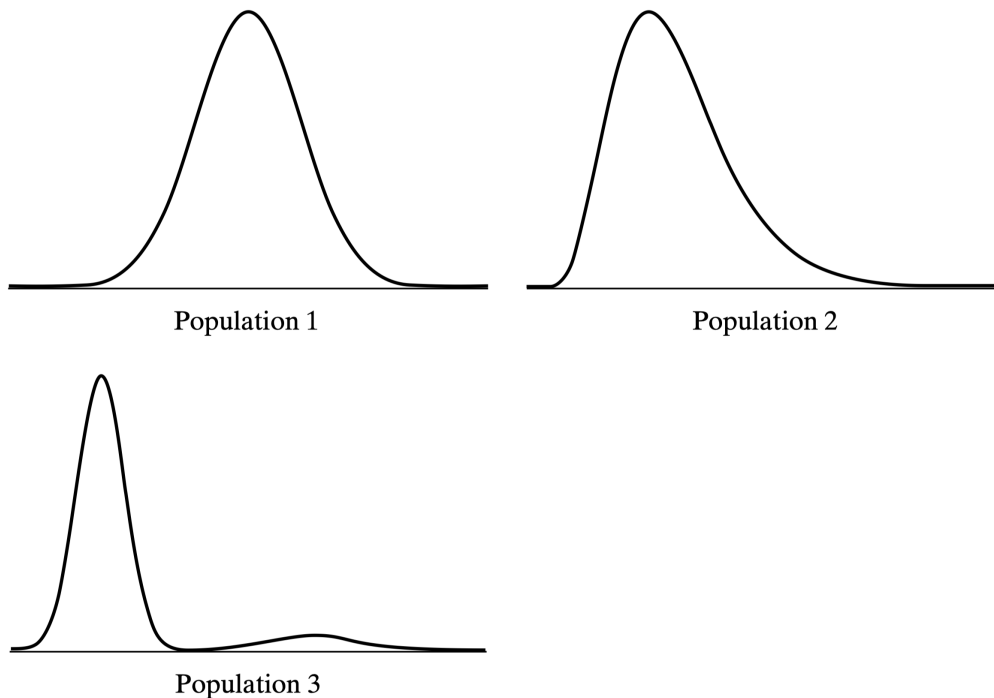
This equation is easily solved to give  $n \geq 38.4$ . Since one cannot have 38.4 butterflies, the new study should include at least 39 butterflies.

- Suppose the researcher in previous example has decided that she would like to be able to estimate the population mean,  $\mu$ , to within  $\pm 0.8$  with 95% confidence.
- That is, she would like her 95% confidence interval for  $\mu$  to be  $\bar{Y} \pm 0.8$ .
- The " $\pm$  part" of the confidence interval, which is called the **margin of error** for 95% confidence, is denoted by  $e = t_{n-1}(0.025) \times SE$ . The precise value of  $t_{n-1}(0.025)$  depends on the degrees of freedom, but typically  $t_{n-1}(0.025)$  is **approximately 2**.
- Thus, the researcher wants  $2 \times SE$  to be no more than 0.8. This means that the SE should be no more than  $0.4 \text{ cm}^2$ .

## Conditions for Validity of a Confidence Interval for $\mu$

- If  $Y$  follows a normal distribution in the population, then Student's  $t$  method is exactly valid. That is to say, the probability that the confidence interval will contain  $\mu$  is actually equal to the confidence level (e.g., 95%).
- By the same token, this interpretation is approximately valid if the population distribution is approximately normal.

- Even if the population distribution is not normal, the Student's  $t$  confidence interval is approximately valid if the sample size is large.
- From a practical point of view, the important question is: How large must the sample be in order for the confidence interval to be approximately valid if the population is nonnormal?
- Not surprisingly, the answer to this question depends on the degree of nonnormality of the population distribution: If the population is only moderately nonnormal, then  $n$  need not be very large. Consider the following three population distributions: (1) normal, (2) slightly skewed right, (3) heavily skewed right.



The table below shows the actual probability that a Student's  $t$  confidence interval will contain  $\mu$  for samples from three different populations.

Table 6.5.2 Actual probability that confidence intervals will contain the population mean							
(a) 95% confidence interval							
	Sample size						
	2	4	8	16	32	64	Very large
Population 1	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Population 2	0.94	0.93	0.94	0.94	0.95	0.95	0.95
Population 3	0.87	0.53	0.57	0.80	0.88	0.92	0.95
(b) 99% confidence interval							
	Sample size						
	2	4	8	16	32	64	Very large
Population 1	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Population 2	0.99	0.98	0.98	0.98	0.99	0.99	0.99
Population 3	0.97	0.82	0.60	0.81	0.93	0.96	0.99

In summary, Student's  $t$  method of constructing a confidence interval for  $m$  is appropriate if the following conditions hold.

- Conditions on the design of the study
  - It must be reasonable to regard the data as a **random sample** from a large population.
  - The observations in the sample must be **independent** of each other.
- Conditions on the form of the population distribution
  - If  $n$  is small, the population distribution must be **approximately normal**.
  - If  $n$  is large, the population distribution need not be approximately normal. In many practical situations, moderate sample sizes (say,  $n = 30$ ) are large enough.

The requirement that the data are a random sample is the most important condition.

## Comparing Two Means

- In previous sections we have considered the analysis of a single sample of quantitative data. In practice, however, much scientific research involves the comparison of two or more samples from different populations. When the observed variable is quantitative, the comparison of two samples can include several aspects, notably (1) **comparison of means**, (2) comparison of standard deviations, and (3) comparison of shapes.
- The notation for comparison of two samples is exactly parallel to our earlier notation, but now a subscript (1 or 2) is used to differentiate between the two samples. The parameter of interest is the difference between two population means  $\mu_1 - \mu_2$ .

### Standard error of $\bar{Y}_1 - \bar{Y}_2$

- To compare two sample means, it is natural to consider the difference between them:  $\bar{Y}_1 - \bar{Y}_2$ , which is an estimate of the quantity  $\mu_1 - \mu_2$ . To characterize the sampling error of estimation, we need to be concerned with the standard error of the difference  $\bar{Y}_1 - \bar{Y}_2$ .
- Recall that  $\text{Var}(Y_1 - Y_2) = \text{Var}(Y_1) + \text{Var}(Y_2)$  if  $Y_1$  and  $Y_2$  are independent. The standard deviation of  $\bar{Y}_1 - \bar{Y}_2$  is thus

$$\sigma_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}}.$$

- Replacing the population SDs  $\sigma_1$  and  $\sigma_2$  with their estimates  $s_1$  and  $s_2$  yields the standard error of  $\bar{Y}_1 - \bar{Y}_2$ :

$$\text{SE}_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}.$$

- When we have two independent samples, we take the SE of each mean, square them, add them, and then take the square root of the sum.

## Example: vital capacity

Vital capacity is a measure of the amount of air that someone can exhale after taking a deep breath. One might expect that musicians who play brass instruments would have greater vital capacities, on average, than would other persons of the same age, sex, and height. In one study the vital capacities of seven brass players were compared to the vital capacities of five control subjects; the table below shows the data.

	Brass player	Control
	4.7	4.2
	4.6	4.7
	4.3	5.1
	4.5	4.7
	5.5	5.0
	4.9	
	5.3	
<i>n</i>	7	5
$\bar{y}$	4.83	4.74
<i>s</i>	0.435	0.351

For the vital capacity data, preliminary computations yield the results in the following table.

	Brass player	Control
$s^2$	0.1892	0.1232
<i>n</i>	7	5
SE	0.164	0.157

The SE of  $\bar{Y}_1 - \bar{Y}_2$  is

$$SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{0.1892}{7} + \frac{0.1232}{5}} = 0.227.$$

Note that

$$0.227 = \sqrt{0.164^2 + 0.157^2}$$

and the SE of the difference is greater than either of the individual SEs but less than their sum.



## Confidence Intervals for $\mu_1 - \mu_2$

One way to compare two sample means is to construct a confidence interval for the difference in the population means. That is, a confidence interval for the quantity  $\mu_1 - \mu_2$ .

Recall that a  $1 - \alpha$  confidence interval for the mean  $\mu$  of a single population that is normally distributed is constructed as

$$\bar{Y} \pm t_{n-1}(\alpha/2) \times SE_{\bar{Y}}.$$

Analogously, a  $1 - \alpha$  confidence interval for  $\mu_1 - \mu_2$  is constructed as

$$(\bar{Y}_1 - \bar{Y}_2) \pm t_{\nu}(\alpha/2) \times SE_{\bar{Y}_1 - \bar{Y}_2},$$

with the degrees of freedom

$$\nu = \frac{(SE_1^2 + SE_2^2)^2}{SE_1^4/(n_1 - 1) + SE_2^4/(n_2 - 1)},$$

where  $SE_1 = s_1/\sqrt{n_1}$  and  $SE_2 = s_2/\sqrt{n_2}$ .

### Example: fast plant

The Wisconsin Fast Plant, *Brassica campestris*, has a very rapid growth cycle that makes it particularly well suited for the study of factors that affect plant growth. In one such study, seven plants were treated with the substance Ancyimidol (ancy) and were compared to eight control plants that were given ordinary water. Heights of all of the plants were measured, in cm, after 14 days of growth. The data are given in the following table. Assume the plant height is normally distributed. Find the 95% confidence interval for  $\mu_1 - \mu_2$ .

Table 6.7.1 Fourteen-day height of control and of ancy plants (cm)		
	Control (Group 1)	Ancy (Group 2)
	10.0	13.2
	13.2	19.5
	19.8	11.0
	19.3	5.8
	21.2	12.8
	13.9	7.1
	20.3	7.7
	9.6	
<i>n</i>	8	7
$\bar{y}$	15.9	11.0
<i>s</i>	4.8	4.7
SE	1.7	1.8

The SE for the difference in sample means is

$$SE_{Y_1 - Y_2} = \sqrt{\frac{4.8^2}{8} + \frac{4.7^2}{7}} = 2.46.$$

Using the formula, we find the degrees of freedom to be:

$$\nu = \frac{(1.7^2 + 1.8^2)^2}{1.7^4/(8-1) + 1.8^4/(7-1)} = 12.8.$$

Using a computer, we can find that for a 95% confidence interval the  $t$  multiplier for 12.8 degrees of freedom is  $t_{12.8}(0.025) = 2.164$ . (Without a computer, we could round down the degrees of freedom, in which case the  $t$  multiplier is  $t_{12}(0.025) = 2.179$ . This change from 12.8 to 12 degrees of freedom has little effect on the final answer.)

The confidence interval formula gives

$$(15.9 - 11.0) \pm 2.164 \times 2.46$$

or

$$4.9 \pm 5.32.$$

The 95% confidence interval for  $\mu_1 - \mu_2$  is

$$(-0.42, 10.22).$$

Thus, we are 95% confident that the population average 14-day height of fast plants when water is used ( $\mu_1$ ) is between 0.42 cm lower and 10.22 cm higher than the average 14-day height of fast plants when ancy is used ( $\mu_2$ ).

## Summary of Estimation Methods

- Standard error of the mean:

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}.$$

- Confidence interval for  $\mu$ :

$$1 - \alpha \text{ confidence interval: } \bar{Y} \pm t_{n-1}(\alpha/2) \times SE_{\bar{Y}}.$$

- The confidence interval formula is valid if (1) the data can be regarded as a **random sample** from a large population, (2) the observations are **independent**, and (3) the population is **normal**. If  $n$  is large then condition (3) is less important.

- Standard error of  $\bar{Y}_1 - \bar{Y}_2$ :

$$SE_{Y_1 - Y_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}} = \sqrt{SE_1^2 + SE_2^2}.$$

- Confidence interval for  $\mu_1 - \mu_2$ :

$$1 - \alpha \text{ confidence interval: } (\bar{Y}_1 - \bar{Y}_2) \pm t_\nu(\alpha/2) \times \text{SE}_{\bar{Y}_1 - \bar{Y}_2}$$

with

$$\nu = \frac{(\text{SE}_1^2 + \text{SE}_2^2)^2}{\text{SE}_1^4/(n_1 - 1) + \text{SE}_2^4/(n_2 - 1)}.$$

- The confidence interval formula is valid if (1) the data can be regarded as coming from two **independently chosen random samples**, (2) the observations are **independent** within each sample, and (3) each of the populations is **normally distributed**. If  $n_1$  and  $n_2$  are large, condition (3) is less important.