Unit 4
Confidence Intervals

Recap

The **sampling distribution** of $\bar{X}$ is the distribution of the sample mean for all possible samples of size n from the population distribution of X.

$$\text{If } X \sim N(\mu, \sigma), \text{ then } \bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

for **any** sample size n.

Central Limit Theorem

If $X \sim$ **?**(μ, σ), then $\bar{X} \dot{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

for **large** values of n (we use $n \geq 30$).

That is, **regardless** of the population distribution of X, the sampling distribution of $\bar{X}$ is approximately normal, provided that the sample size is large.

Statistical Inference

We get data from a sample, but we are often not satisfied with information just about the sample itself. We would like to use the sample data to **infer** something about the population of interest.

Statistical Inference provides methods for drawing conclusions about a population from sample data.

Statistical Inference

We can never be sure that our sample data fairly represent the population. In order to quantify this uncertainty, in statistical inference we use the language of **probability**.

Like probability, the foundation of inference lies on long-run predictable behaviour.

By taking "good" samples (i.e., SRS), we can draw conclusions with a high probability of being correct.

Statistical Inference

We have used the sample mean $\bar{X}$ as an estimator for the population mean μ, but we have never asked the question:

How good of an estimator is $\bar{X}$?

The probability that $\bar{X} = \mu$ is equal to zero because of continuity. Reporting the sample mean alone gives us **no information** as to **how accurate** we believe our estimate to be.

Confidence Intervals

We would like to construct an **interval** of values to estimate the population mean.

We know the sample mean will vary from sample to sample. Suppose we were to take many samples of the same size n.

We would like to construct the interval in such a way that μ is contained in the interval for **most samples**. That is, we would like to be **confident** that the interval we construct contains the value of the parameter we are trying to estimate.

95% Confidence Interval

We use the sample mean to estimate μ, so it is logical that we should center our interval at $\bar{X}$.

Recall the 68-95-99.7 rule tells us that when a random variable follows a normal distribution, about 95% of all values of that variable fall within two standard deviations of its mean.

So if $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, about 95% of all values of $\bar{X}$ fall

within $2\sigma/\sqrt{n}$ of the mean μ.

95% Confidence Interval

If $\bar{X}$ is within two standard deviations of μ, it follows that μ is within two standard deviations of $\bar{X}$. This happens in about 95% of all samples.

That is, in 95% of all samples, μ lies within

$$\bar{x} \pm 2\frac{\sigma}{\sqrt{n}}$$

This interval of values is called the **95% confidence interval for μ**.

Example

Suppose it is known that GPAs of University of Manitoba graduates follow a normal distribution with standard deviation $\sigma = 0.40$.

A simple random sample of 25 graduates is selected and their mean GPA is calculated to be 3.31.

Example

We know that

$$\bar{x} = 3.31,\ \sigma = 0.40,\ n = 25$$

A 95% confidence interval for the true mean GPA of all U of M graduates is:

$$\bar{x} \pm 2\frac{\sigma}{\sqrt{n}} = 3.31 \pm 2\left(\frac{0.40}{\sqrt{25}}\right)$$

$$= 3.31 \pm 0.16$$

$$= (3.15,\ 3.47)$$

Confidence Intervals

Note that there are only two possibilities here:

- The true value of μ actually falls within this interval.

- This is one of the rare samples (5%) that produces an interval which excludes the true value of μ.

Confidence Interval Interpretation

We interpret the 95% confidence interval for μ as follows:

"If we repeatedly took simple random samples of the same size from the same population and constructed the interval in a similar manner, 95% of all such intervals would contain the true mean μ of the population."

Confidence Intervals

Like all confidence intervals we will encounter in this course, this one has the form:

$$\text{estimate} \pm \text{margin of error}$$

The **estimate** is our best guess at the true value of μ, while the **margin of error** reflects how accurate we believe our estimate to be.

For the 95% confidence interval, the estimate is $\bar{x}$ and the margin of error is $2\sigma/\sqrt{n}$.

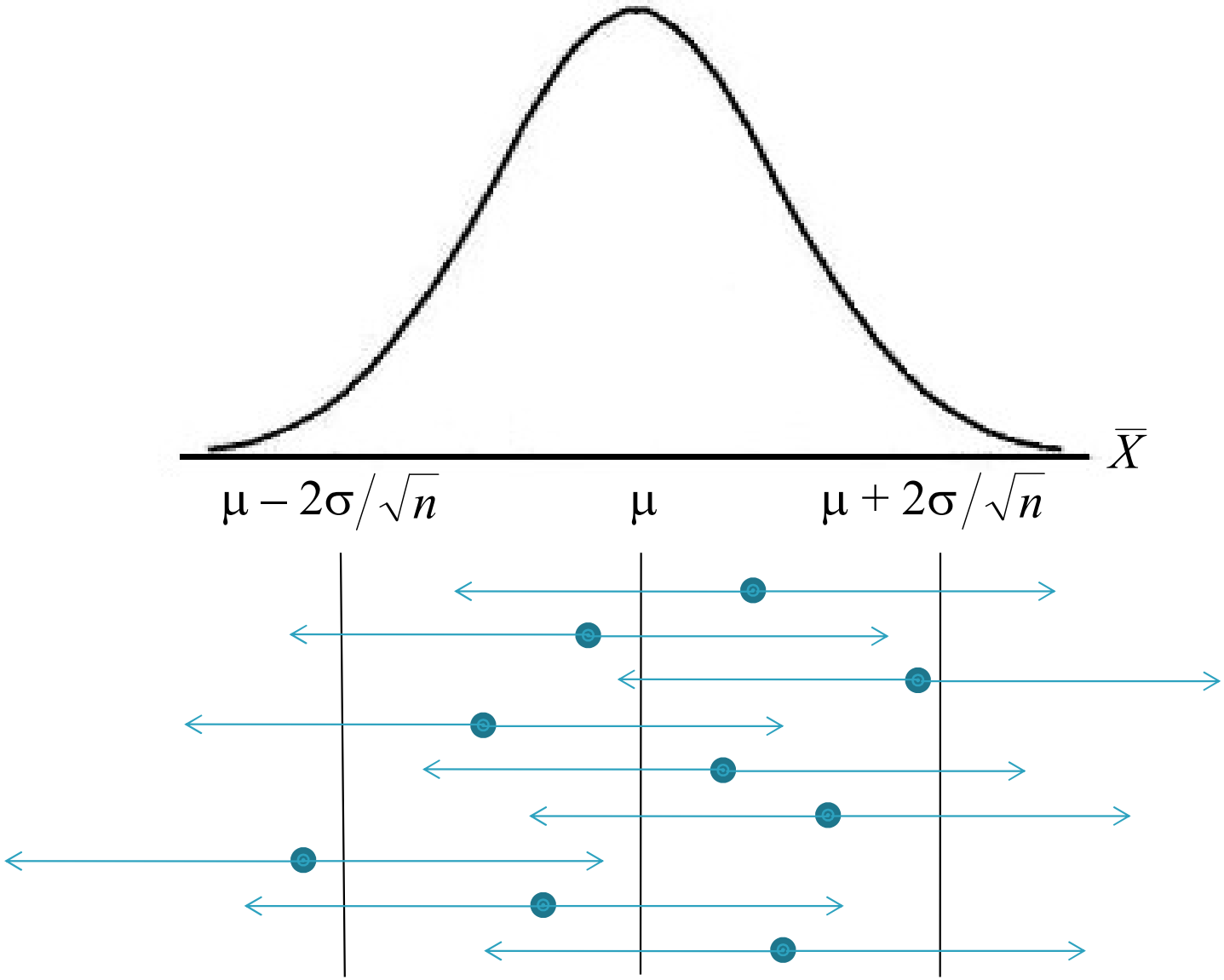
Confidence Level

Each confidence interval has associated with it a **confidence level C**, which gives the probability that the interval will capture the true value of the population mean μ.

Of course, there is no reason that we have to use a 95% confidence level.

We choose the confidence level ourselves. We always use a **high** confidence level (usually 90% or higher), as our goal is to estimate the parameter with a **high** probability of accuracy.

A picture of what we are doing:



Confidence Intervals

Note that all points falling within two standard deviations of μ (we know this to be about 95%) catch the value of the population mean when stretched on either side by $2\sigma/\sqrt{n}$.

Confidence Intervals

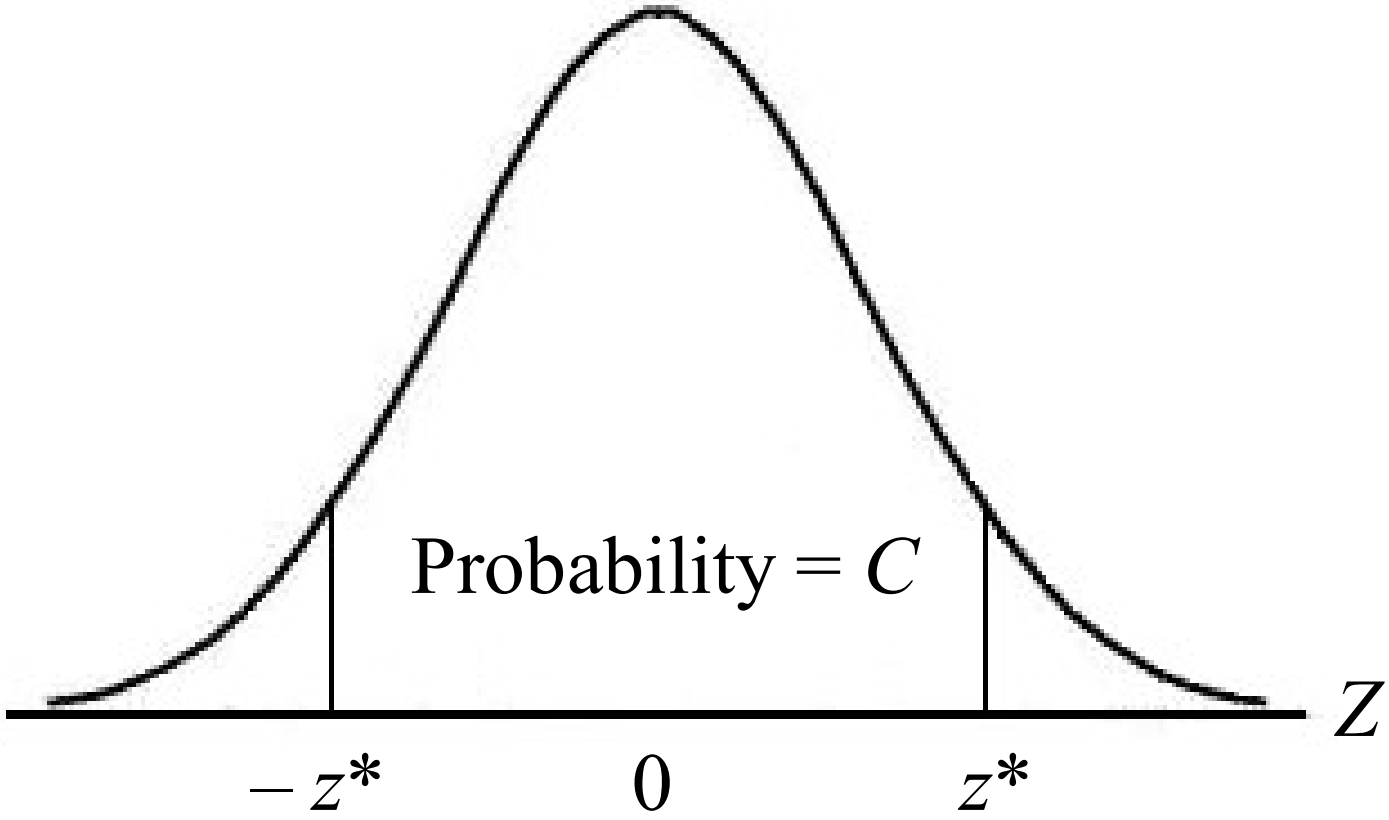
The form of a general level C confidence interval for the population mean μ is:

$$\bar{x} \pm z^* \left(\frac{\sigma}{\sqrt{n}} \right)$$

where z^* is the value of Z such that

$$P(-z^* \leq Z \leq z^*) = C$$

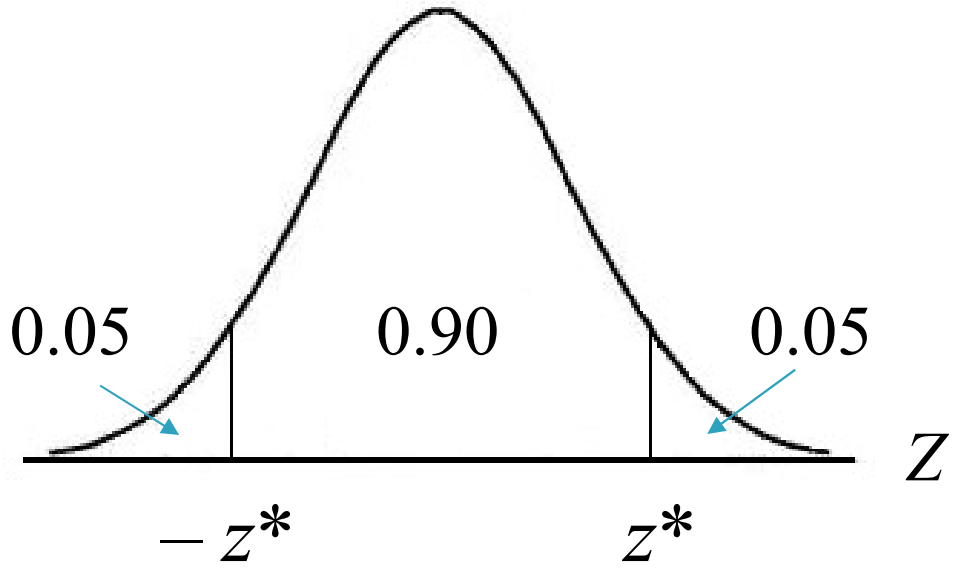
Confidence Intervals



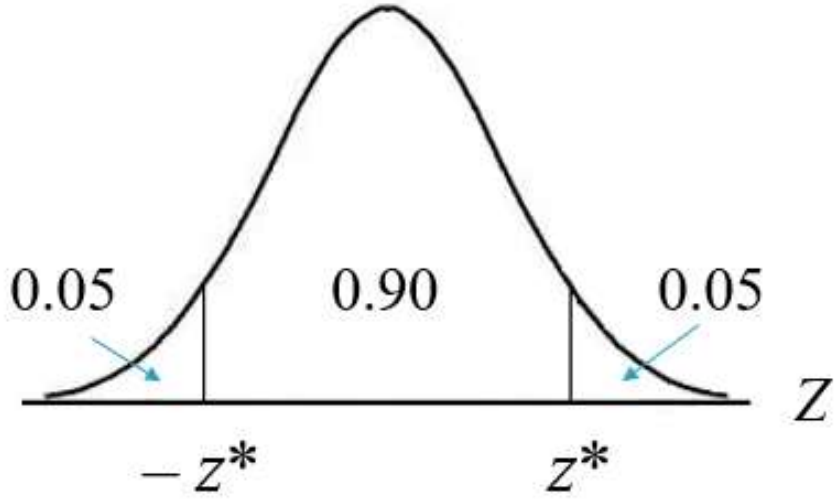
Confidence Intervals

For example, to find a 90% confidence interval, we must catch the central 90% of the normal distribution.

In catching the central 90%, we must leave out 10%, or 5% in each tail of the distribution.



Confidence Intervals



We find the value z^* such that

$$P(Z \leq z^*) = 0.90 + 0.05 = 0.95$$

We consult Table 1 and find that $z^* = 1.645$.

Critical Values

You can find z^* for any confidence level C by searching Table 1. For example, the value z^* corresponding to a 95% confidence interval is actually 1.96, and not 2, as we have been using as an approximation.

The values of z^* for the most common confidence levels (90%, 95%, 99%, etc.) are given in the last line of Table 2.

Values z^* that mark off a specific area under the standard normal curve are called **critical values** of the distribution.

Example

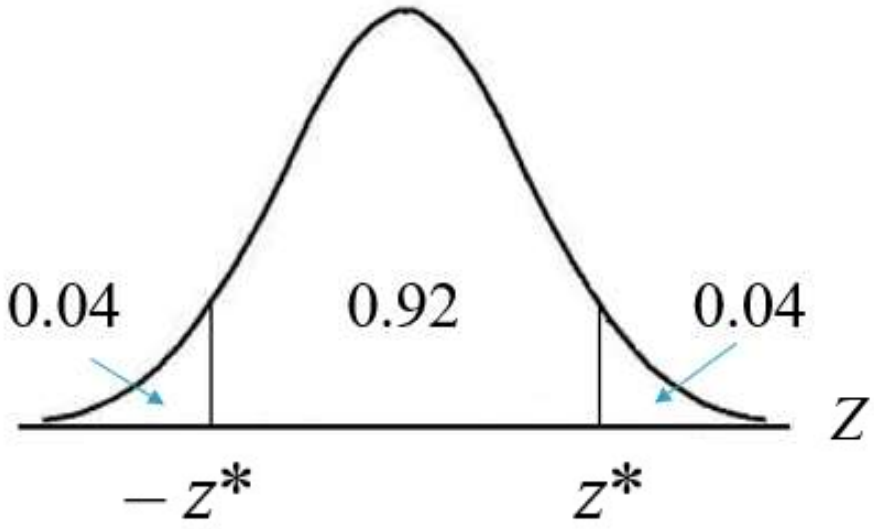
The sentence times for criminals convicted of a particular crime follow a normal distribution with standard deviation $\sigma = 28.7$ months. The sentences (in months) of a random sample of ten criminals convicted of this crime are shown below:

136	102	84	150	115
98	125	176	120	72

Example

Suppose we wish to find a 92% confidence interval for μ:

We calculate $\bar{x} = 117.8$ and we find the value z^* such that $P(Z \leq z^*) = 0.92 + 0.04 = 0.96$. This corresponds to the critical value $z^* = 1.75$.



Example

Our 92% confidence interval is thus:

$$\bar{x} \pm z^* \left(\frac{\sigma}{\sqrt{n}}\right) = 117.8 \pm 1.75 \left(\frac{28.7}{\sqrt{10}}\right)$$

$$= 117.8 \pm 15.9 = (101.9, 133.7)$$

We are 92% confident that this interval contains the true mean sentence of all criminals convicted of this crime. That is, **if we collected many samples of the same size and computed the interval in a similar manner, 92% of such intervals would contain the value of the population mean μ**.

R Code

```
> Sentence <- c(136, 102, 84, 150, 115, 98, 125,
                176, 120, 72)
> install.packages("BSDA")
> library(BSDA)
> z.test(Sentence, sigma.x = 28.7, conf.level = 0.92)

        One-sample z-Test

data:  Sentence
z = 12.98, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
92 percent confidence interval:
 101.9112 133.6888
sample estimates:
mean of x
    117.8
```

Example

Suppose instead we wanted a 99% confidence interval for μ. The only difference in our formula for the confidence interval is the value of z^*. We can take the value from Table 2. It is $z^* = 2.576$.

So our 99% confidence interval is:

$$\bar{x} \pm z^* \left(\frac{\sigma}{\sqrt{n}} \right) = 117.8 \pm 2.576 \left(\frac{28.7}{\sqrt{10}} \right)$$

$$= 117.8 \pm 23.4 = (94.4, 141.2)$$

Example

92% C.I.: (101.9, 133.7)
99% C.I.: (94.4, 141.2)

Notice that as the confidence level increases, so too does the margin of error (and hence the length of the interval).

If we increase the confidence level, we must sacrifice our precision of estimation. If we want to be **more sure** that our interval contains μ, we have to **expand** the interval!

Practice Question

An airport manager would like to estimate the true mean number of minutes μ passengers arrive before their scheduled flight departure. Suppose it is known that arrival times follow a normal distribution with standard deviation 15 minutes. A random sample of 9 passengers arrived an average of 70 minutes before their flight. A 98% confidence interval for μ is:

(A) (58.37, 81.63) (B) (59.73, 80.27) (C) (65.10, 74.90)

(D) (57.62, 82.38) (E) (61.45, 78.55)

Practice Question

Lengths of pet goldfish are known to follow a normal distribution with standard deviation 0.38 cm. A random sample of 20 goldfish is selected and their mean length is calculated to be 3.21 cm. An 85% confidence interval for the true mean length of all pet goldfish is:

(A) $3.21 \pm 1.53\left(\frac{0.38}{\sqrt{20}}\right)$

(B) $3.21 \pm 1.44\left(\frac{0.38}{\sqrt{20}}\right)$

(C) $3.21 \pm 1.04\left(\frac{0.38}{\sqrt{20}}\right)$

(D) $3.21 \pm 0.85\left(\frac{0.38}{\sqrt{20}}\right)$

(E) $3.21 \pm 1.46\left(\frac{0.38}{\sqrt{20}}\right)$

Example

We would like to estimate the true mean hourly wage of all employees of a large national company with 95% confidence. Suppose it is known that the population standard deviation is σ = \$11.23 per hour.

A random sample of 40 employees gives a mean of \$21.74 per hour.

Example

Our 95% confidence interval for μ is:

$$\bar{x} \pm z^* \left(\frac{\sigma}{\sqrt{n}} \right) = 21.74 \pm 1.96 \left(\frac{11.23}{\sqrt{40}} \right)$$

$$= 21.74 \pm 3.48 = (18.26, 25.22)$$

We are 95% confident that the true mean hourly wage for all employees of this company is between $18.26 and $25.22 per hour. That is, if we were to take repeated samples of 40 employees and compute the interval in a similar manner, then 95% of such intervals would contain the true mean hourly wage.

Example

Notice that there was no mention of wages following a normal distribution. In fact, this is likely not the case.

However, the use of the normal distribution in this case is justified because our sample size is sufficiently high to apply the Central Limit Theorem. The 95% confidence level is therefore approximate.

Confidence Intervals

We would ideally like to use a **high** confidence level and obtain a **narrow** confidence interval, but we have seen that there is a trade-off between the confidence level and the margin of error.

How can we reduce the length of the interval without sacrificing our precision of estimation?

Example

In our previous example, suppose we had instead selected a sample of 160 employees and that we had calculated the same sample mean hourly wage of $21.74.

A 95% confidence interval for μ is then:

$$\bar{x} \pm z^* \left(\frac{\sigma}{\sqrt{n}} \right) = 21.74 \pm 1.96 \left(\frac{11.23}{\sqrt{160}} \right)$$

$$= 21.74 \pm 1.74 = (20.00, 23.48)$$

Example

95% C.I. when $n = 40$: ($18.26, $25.22)

95% C.I. when $n = 160$: ($20.00, $23.48)

Notice that a **higher** sample size results in a **lower** margin of error (and hence a **narrower** confidence interval). In fact, we see that taking a sample size that is **four times greater** results in a margin of error only **half** as large.

Effect of Increasing Sample Size

Increasing the sample size by a factor of k reduces the margin of error by a factor of $\sqrt{k}$.

We can always reduce the margin of error by increasing our sample size. As is always the case in statistics, a higher sample size leads to more accurate results.

The interpretation given for the confidence interval in the previous example is the **only** correct interpretation.

Some common misinterpretations

"Approximately 95% of all employees earn between $20.00 and $23.48 per hour."

The interval estimates the **mean** wage and does not apply to individual observations.

Some common misinterpretations

"The probability that the calculated interval contains the sample mean is 95%."

The sample mean $\bar{x}$ is **always** contained in the confidence interval (the interval is centered at $\bar{x}$).

Some common misinterpretations

"About 95% of all samples of 160 employees have means between $20.00 and $23.48 per hour."

The interval estimates the **population mean** μ and does not apply to other potential samples.

Some common misinterpretations

"The probability that μ is between \$20.00 and \$23.48 is 0.95."

The population mean has a **fixed value**. We must be careful to make our probability statements in terms of the interval (which **is** random) rather than in terms of the parameter.

Practice Question

The manager at a grocery store would like to estimate the true mean amount of money spent by customers in the express lane. She selects a simple random sample of 50 receipts and calculates a 97% confidence interval for μ to be ($15.50, $20.25). The confidence interval can be interpreted to mean that, in the long run,

(A) 97% of similarly constructed intervals would contain the population mean.
(B) 97% of similarly constructed intervals would contain the sample mean.
(C) 97% of all customers in the express lane spend between $15.50 and $20.25.
(D) 97% of samples of 50 customers will have means between $15.50 and $20.25.
(E) 97% of customers who are spending between $15.50 and $20.25 use the express lane.

Practice Question

We would like to construct a confidence interval to estimate the mean μ of some variable X. Which of the following combinations of confidence level and sample size will produce the widest interval?

(A) 90% confidence, $n = 10$

(B) 95% confidence, $n = 10$

(C) 95% confidence, $n = 20$

(D) 99% confidence, $n = 10$

(E) 99% confidence, $n = 20$

Practice Question

Prices of hardcover books at a large bookstore are known to follow a normal distribution with standard deviation $12.50. You take a random sample of 17 hardcover books and calculate the sample mean to be $38.76. What is the margin of error for a 99% confidence interval for the true mean price of all hardcover books at the bookstore?

(A) 6.49 (B) 6.73 (C) 7.24 (D) 7.56 (E) 7.81

Sample Size vs. Population Size

Note that the length of a confidence interval depends on the sample size, **but not the population size**!

For a given sample size, our estimate will be **just as precise**, regardless of the size of the population (as long as the population size is reasonably **large**).

Sample Size vs. Population Size

Suppose we wanted to estimate the mean GPA μ for all University of Winnipeg graduates (analogous to our previous example for U of M graduates).

Suppose we take a sample of 25 students from the U of W (which has a population of 10,000 students), the same sample size we took for the U of M (which has a population of 30,000 students).

Sample Size vs. Population Size

Assuming equal standard deviations, a 95% confidence interval for μ for the U of W will have the same margin of error as that for the U of M, despite the fact that the U of M population is three times greater!

Sample Size Determination

When collecting a sample, it is always wise to consider the purpose of our data collection. We would often like to achieve a certain precision of estimation. To accomplish this, we require a sufficient sample size.

Sample Size Determination

Ideally, we would like to have a **small** margin of error, together with a **high** confidence level (i.e., we want a narrow interval which contains the true value of μ with a high probability).

This can all be achieved by choosing the appropriate sample size.

Sample Size Determination

Let m denote the margin of error:

$$m = z^* \frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow \quad n = \left(\frac{z^* \sigma}{m} \right)^2$$

Example

Suppose it is known that the amount of time adults sleep at night follows a normal distribution with standard deviation 1.27 hours.

We would like to take a sample of people large enough to estimate the true mean time adults sleep at night to within 0.5 hours with 98% confidence. How many people do we need to sample in order to achieve this?

Example

We know that

$$z^* = 2.326, \quad \sigma = 1.27, \quad m = 0.5$$

Therefore,

$$n = \left(\frac{z^*\sigma}{m}\right)^2 = \left(\frac{2.326(1.27)}{0.5}\right)^2 = 34.9 \approx 35$$

R Code

```
> nsize(0.5, sigma = 1.27, conf.level = 0.98)
```

```
The required sample size (n) to estimate the population
mean with a 0.98 confidence interval so that the margin
of error is no more than 0.5 is 35 .
```

Effect of Decreasing Margin of Error

We need a minimum sample of 35 people. Note that we always round up, as we calculated that we need a sample of **at least** 34.9 people. (We would even round the value 34.01 up to 35.)

Now suppose that we decide that a margin of error of 0.5 hours is too large, and we would like to estimate the true mean time adults sleep at night to within 0.25 hours with 98% confidence (i.e., cut the margin of error in half).

Effect of Decreasing Margin of Error

We require a sample of size

$$n = \left(\frac{z^*\sigma}{m}\right)^2 = \left(\frac{2.326(1.27)}{0.25}\right)^2 = 139.62 \approx 140$$

Notice that when we cut the margin of error in half, we require four times the sample size. In general, **if we want to reduce the margin of error by a factor of *k*, we need a sample that is k^2 times as large**. If we want to reduce the margin of error to one third its original value, we need nine times more individuals in our sample, etc.

Practice Question

We would like to estimate the true mean height of all male Olympic swimmers. Suppose that heights are known to follow a normal distribution with standard deviation 4.4 cm. What sample size is required in order to estimate the true mean to within 1.5 cm with 96% confidence?

(A) 19 (B) 24 (C) 37 (D) 48 (E) 53

Practice Question

A real estate agent would like to estimate the true mean value of all houses in Winnipeg. She calculates that, in order to estimate the true mean to within $10,000 with 95% confidence, she needs to select a sample of 90 houses in Winnipeg.

What sample size would be required to estimate the true mean value of all houses in Winnipeg to within $2,500 with 95% confidence?

(A) 6 (B) 23 (C) 180 (D) 360 (E) 1440

Practice Question

A real estate agent would like to estimate the true mean value of all houses in Winnipeg. She calculates that, in order to estimate the true mean to within $10,000 with 95% confidence, she needs to select a sample of 90 houses in Winnipeg.

The city of Saskatoon has only one third as many houses as Winnipeg. Assuming the standard deviation of house values is the same in both cities, what sample size would be required to estimate the true mean value of all houses in Saskatoon to within $10,000 with 95% confidence?

(A) 10 (B) 30 (C) 90 (D) 270 (E) 810

Some cautions

- Our formula for the confidence interval holds only if the data were collected using an SRS. There is no correct way to do proper inference using data collected haphazardly. Good formulas cannot rescue us from poor sampling methods.

- Since the sample mean is strongly influenced by outliers, so too is the confidence interval.

Some cautions

- We are using the true population standard deviation σ in our calculations. In practice, this is not a realistic assumption. We will see a proper method for constructing confidence intervals when we only have the sample standard deviation s. We make this unreasonable assumption now to establish the framework for building confidence intervals.

- The margin of error covers **only** random sampling errors. It does not reflect any degree of undercoverage, nonresponse, or other forms of bias.

Practice Question

Summer temperatures in Las Vegas, Nevada are known to follow a normal distribution with standard deviation 5.5°C. We take a random sample of 16 summer days in Las Vegas and calculate a sample mean temperature of 39.2°C.

From the data, a 99% confidence interval for the true mean summer temperature in Las Vegas is calculated to be (35.7, 42.7).

Write the correct interpretation for this confidence interval.