Introduction to Inference

Estimating with Confidence

Chapter 8

Data

- Mean time spent online by adults >18 6.15 hours
- 28% of Americans have no savings
- Mean amount of student debt for California student $21,382 (bachelors degree)
- 55% of students in California will graduate with debt
- For U.S. men, the average life expectancy is 76, while it's 81 for U.S. women

Overview of Inference

- Methods for drawing conclusions about a population from sample data are called statistical inference

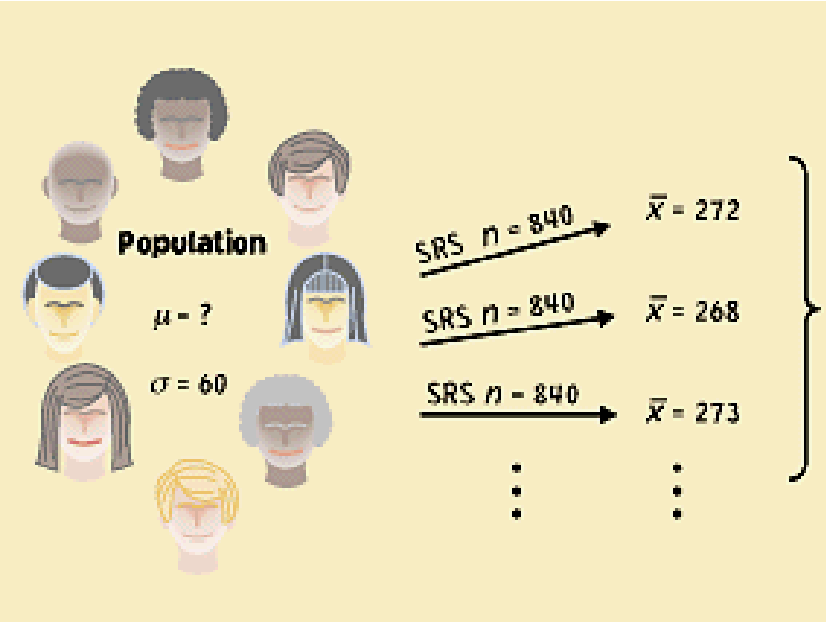
- Methods
  - Confidence Intervals - estimating a value of a population parameter
  - Tests of significance - assess evidence for a claim about a population

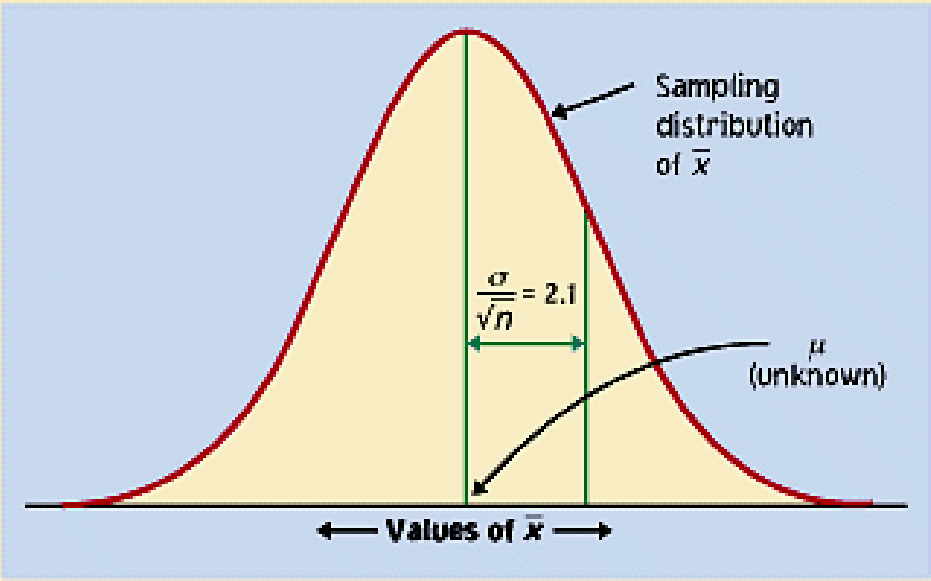
- Inference is appropriate when data are produced by either
  - a random sample or
  - a randomized experiment

Statistical confidence

Although the sample mean, $\bar{x}$, is a unique number for any particular sample, if you pick a different sample you will probably get a different sample mean.

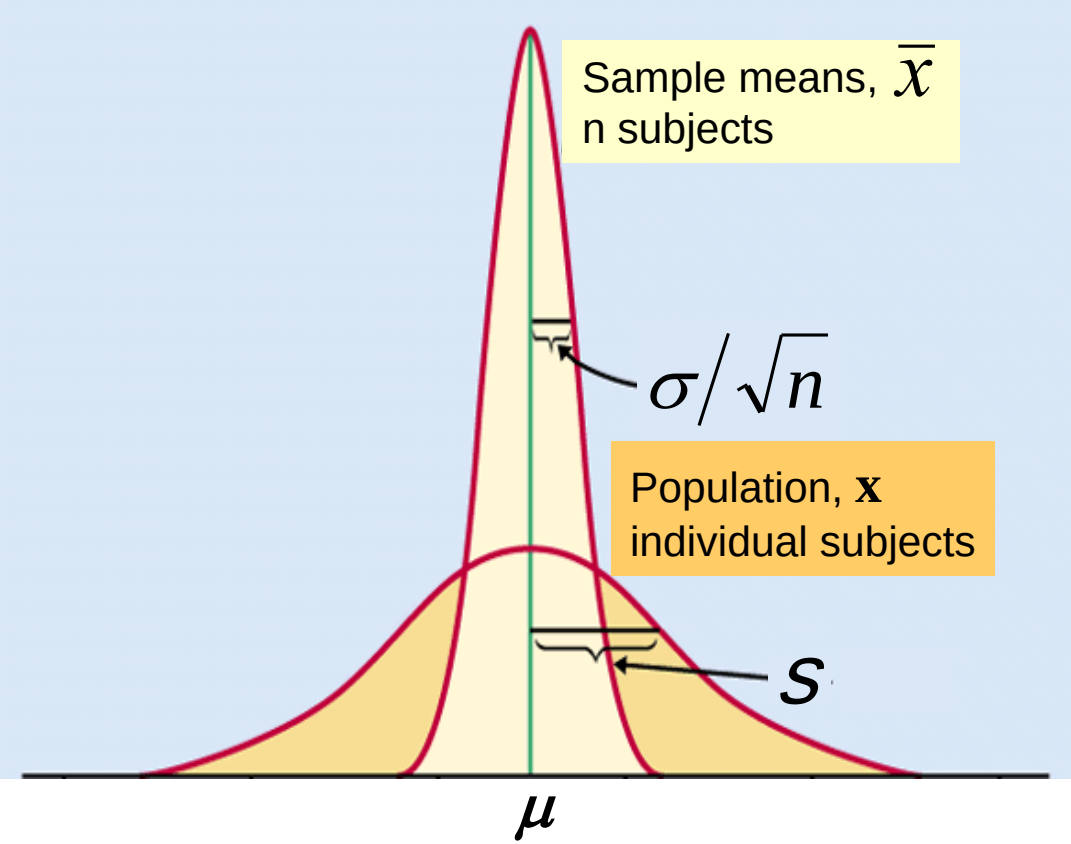
In fact, you could get many different values for the sample mean, and virtually none of them would actually equal the true population mean, μ.





But the sample distribution is narrower than the population distribution, by a factor of $\sqrt{n}$.

Thus, the estimates $\bar{x}$ gained from our samples are always relatively close to the population parameter μ.

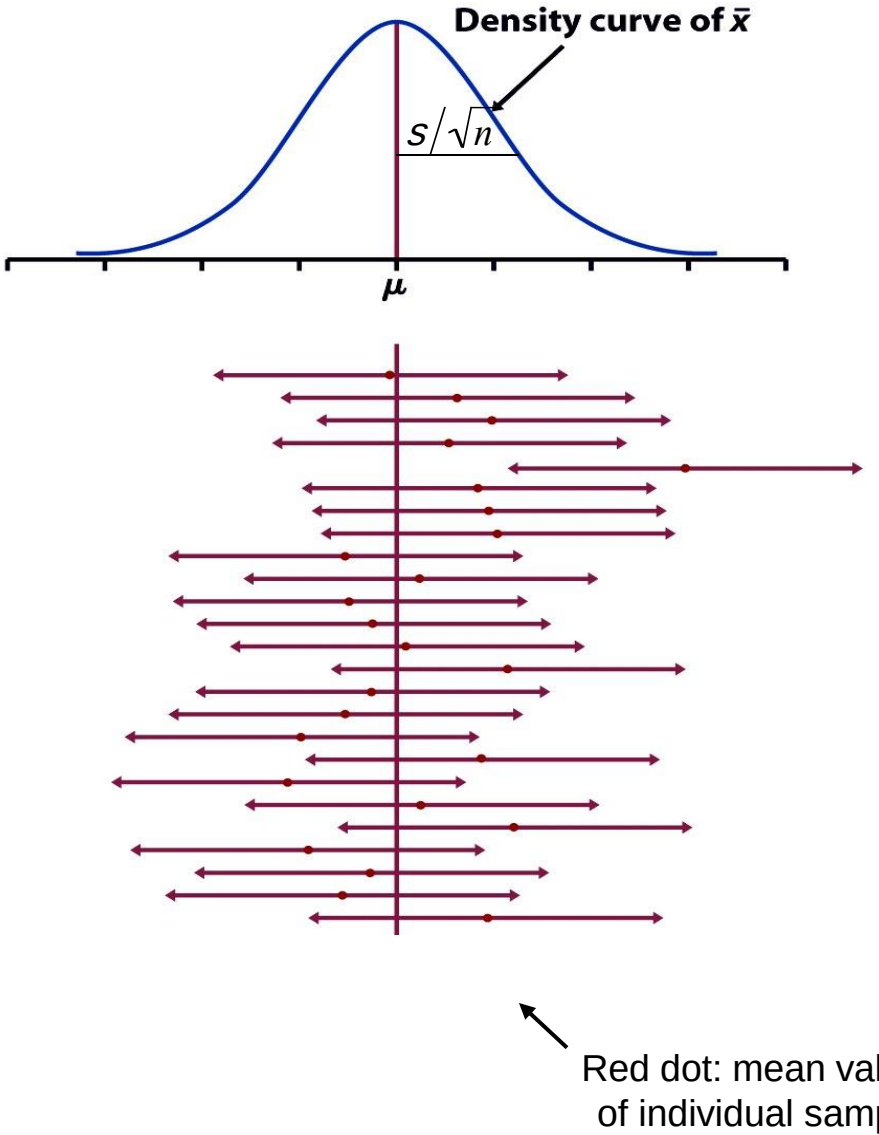


If the population is normally distributed so will the sampling distribution

Remember:

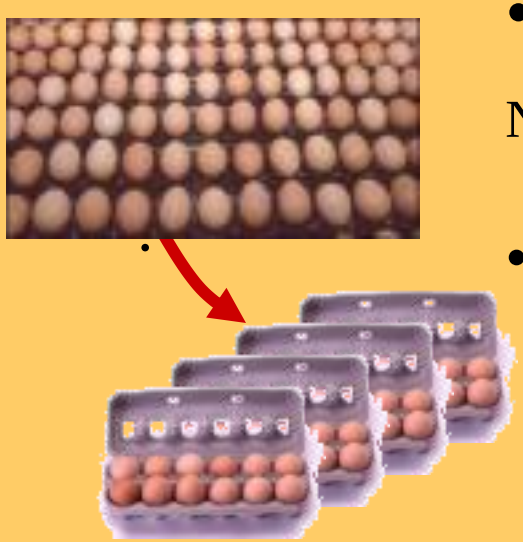
95% of all sample means will be within roughly 2 standard deviations ($2*\sigma/\sqrt{n}$) of the population parameter μ.

Distances are symmetrical which implies that **the population parameter μ must be within roughly 2 standard deviations from the sample average $\bar{x}$, in 95% of all samples.**



This reasoning is the essence of statistical inference.

The weight of single eggs of the brown variety is normally distributed $N(65\text{ g},5\text{ g})$. Think of a carton of 12 brown eggs as an SRS of size 12.



- What is the distribution of the sample means $\bar{x}$?

Normal (mean μ, standard deviation $\sigma/\sqrt{n}$) = $N(65\text{ g},1.44\text{ g})$.

- Find the middle 95% of the sample means distribution.

Roughly $\pm$ 2 standard deviations from the mean, or

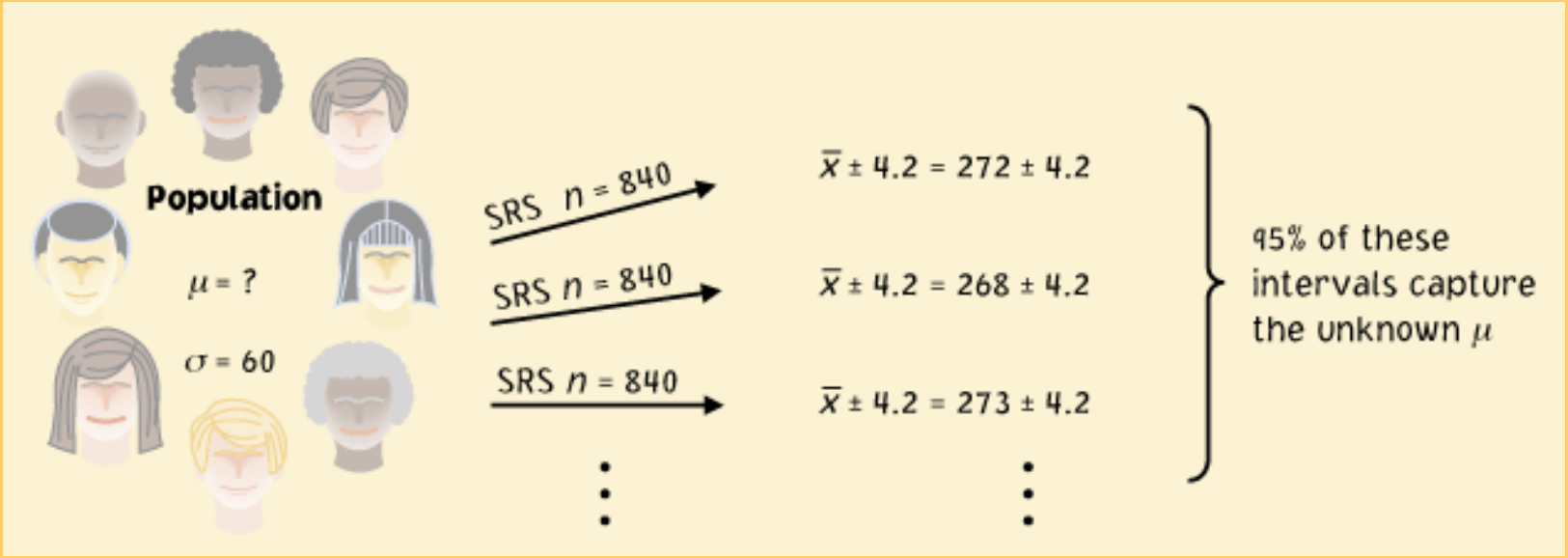
65g $\pm$ 2.88g.

What can you infer about the mean μ of the brown egg population?

There is a 95% chance that the population mean μ is roughly within $\pm\ 2\sigma/\sqrt{n}$ or 65 g $\pm$ 2.88 g.

Confidence intervals

The **confidence interval** is a range of values with an associated probability or **confidence level *C*.** The probability quantifies the chance that the interval contains the true population parameter.

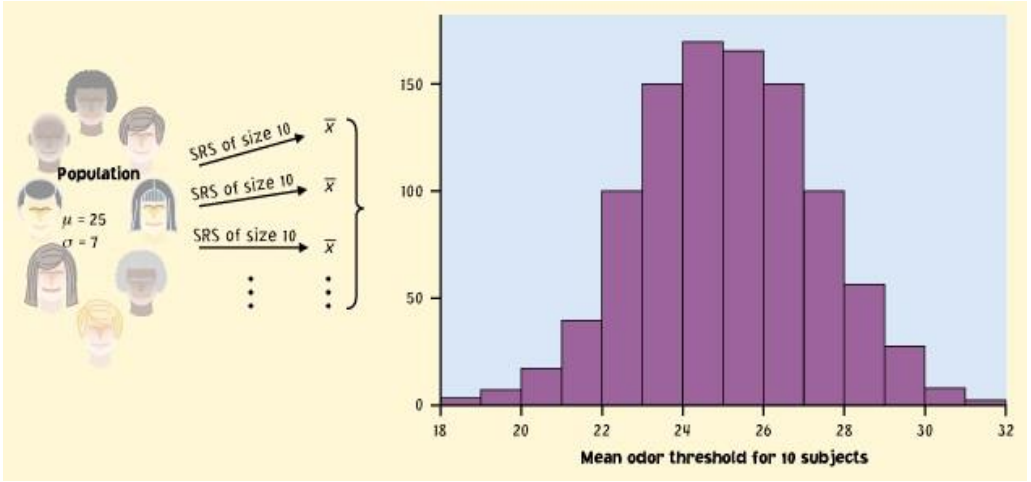


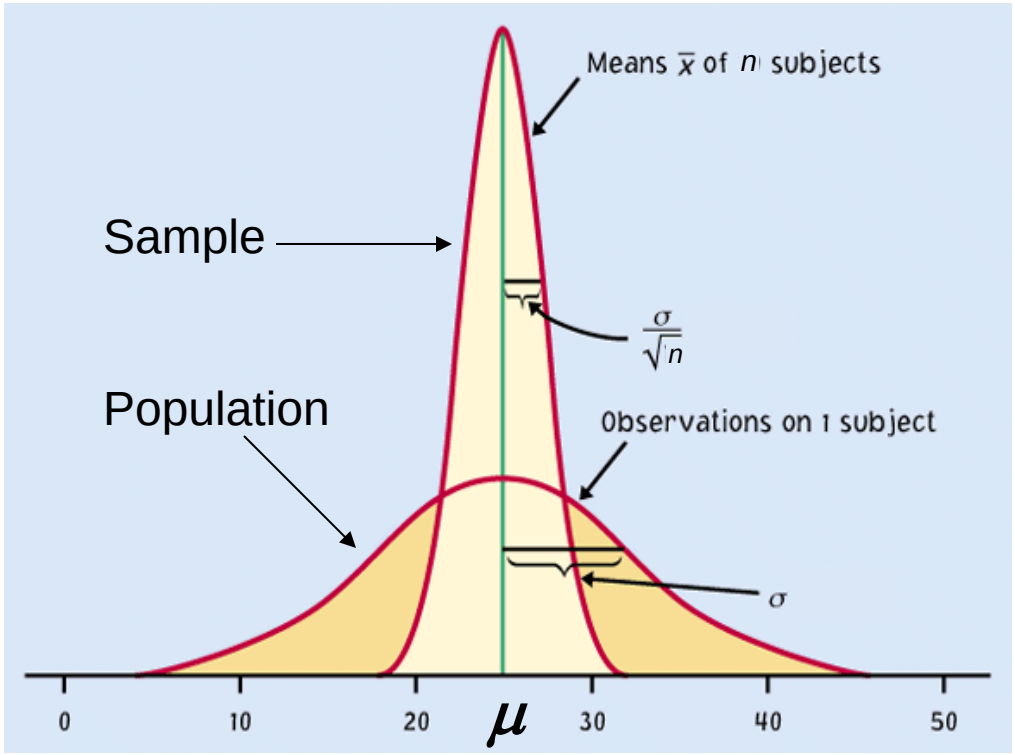
$\bar{x} \pm 4.2$ is a 95% confidence interval for the population parameter μ.

This equation says that in 95% of the cases, the actual value of μ will be within 4.2 units of the value of $\bar{x}$.

Implications

We don't need to take a lot of random samples to "rebuild" the sampling distribution and find μ at its center.

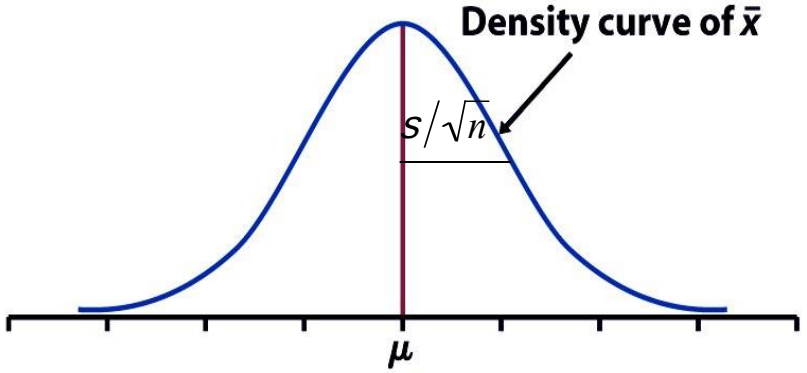


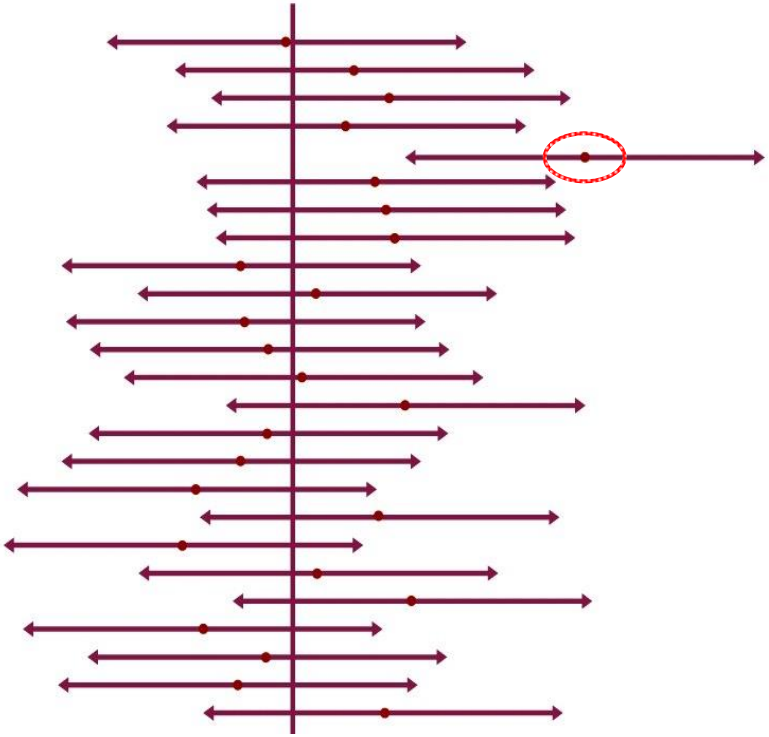


All we need is one SRS of size *n* and rely on the properties of the sample means distribution to infer the population mean μ.

Reworded

With 95% confidence, we can say that μ should be within roughly 2 standard deviations ($2*\sigma/\sqrt{n}$) from our sample mean $\bar{x}$.





- In 95% of all possible samples of this size n, μ will indeed fall in our confidence interval.
- In only 5% of samples would $\bar{x}$ be farther from μ.

A **confidence interval** can be expressed as:

- Mean $\pm E$

  E is called the **margin of error**

  μ within $\bar{x} \pm E$

  Example: 120 ± 6

- Two endpoints of an interval

  μ within $(\bar{x} - E)$ to $(\bar{x} + E)$

  ex. 114 to 126

A **confidence level *C*** (in %)

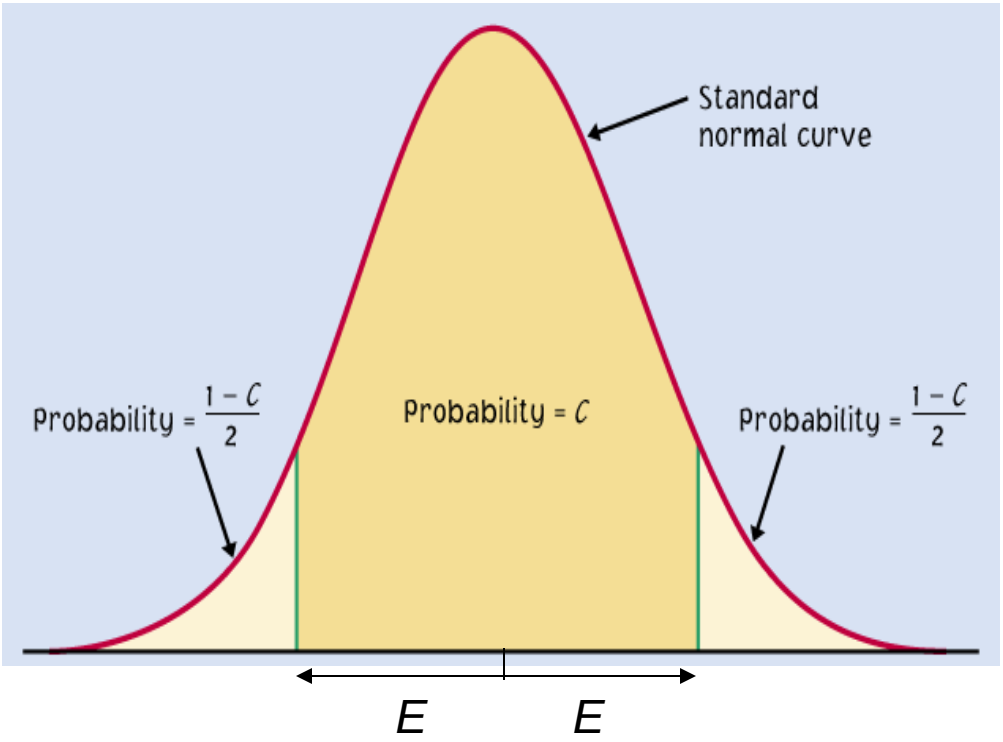
indicates the probability that the

μ falls within the interval.

It represents the area under the

normal curve within $\pm E$ of the

center of the curve.



Example: Point Estimate for Population μ

An economics researcher is collecting data about grocery store employees in a county. The data listed below represents a random sample of the number of hours worked by 40 employees from several grocery stores in the county. Find a point estimate of the population mean, μ.

30	26	33	26	26	33	31	31	21	37
27	20	34	35	30	24	38	34	39	31
22	30	23	23	31	44	31	33	33	26
27	28	25	35	23	32	29	31	25	27

Solution: Point Estimate for Population μ

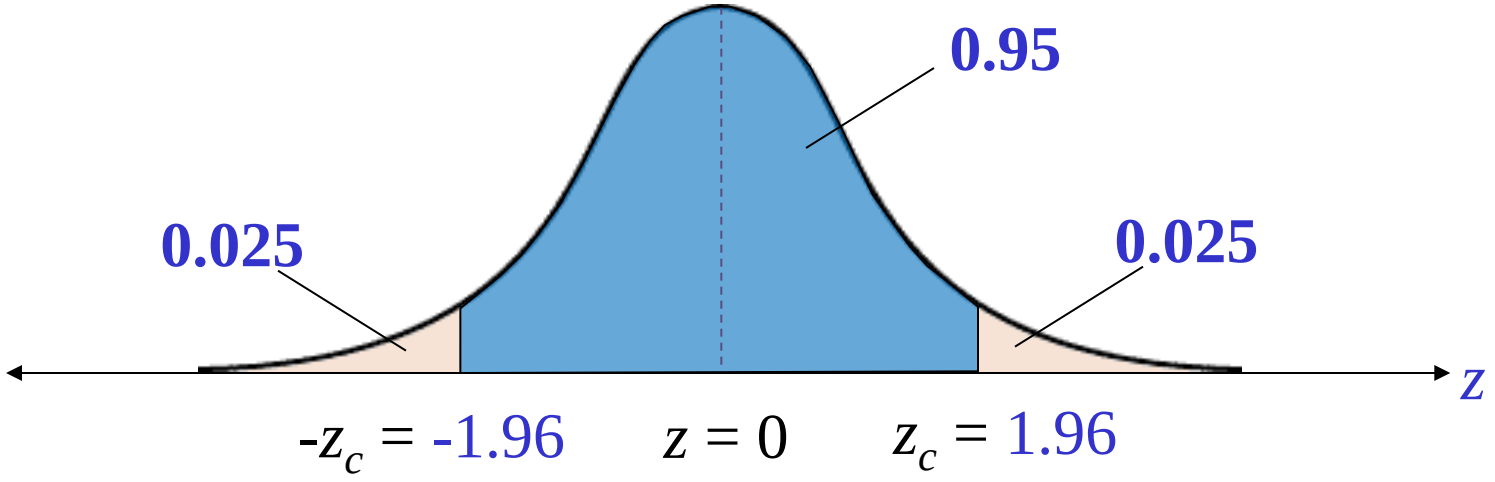
The sample mean of the data is

$$\bar{x} = \frac{\Sigma x}{n} = \frac{1184}{40} = 29.6$$

The point estimate for the mean number of hours worked by grocery store employees in this county is 29.6 hours.

Use the data about the grocery store employees and a 95% confidence level to find the margin of error for the mean number of hours worked by grocery store employees. Assume the population standard deviation is 7.9 hours.

- First find the critical values



95% of the area under the standard normal curve falls within 1.96 standard deviations of the mean. (You can approximate the distribution of the sample means with a normal curve by the Central Limit Theorem, because $n = 40 \geq 30$.)

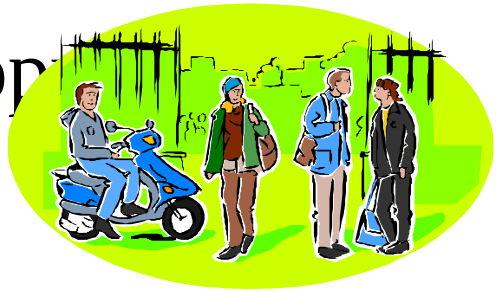
$$E = z_c \frac{s}{\sqrt{n}}$$

$$\approx 1.96 \times \frac{7.9}{\sqrt{40}}$$

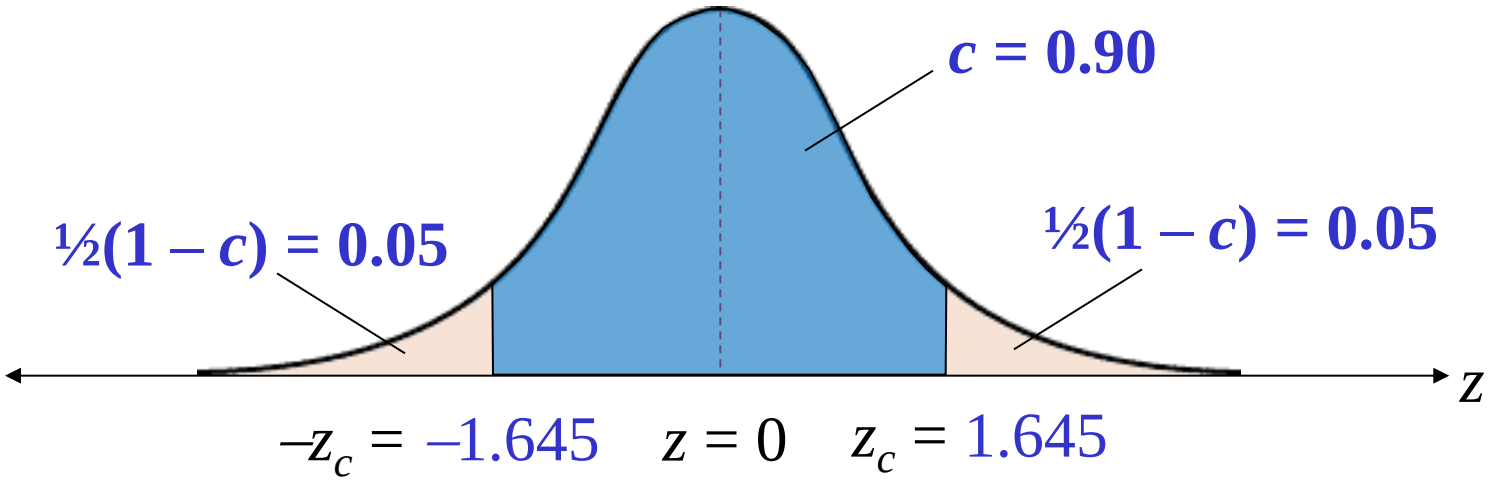
$$\approx 2.4$$

You are 95% confident that the margin of error for the population mean is about 2.4 hours.

A college admissions director wishes to estimate the mean age of all students currently enrolled. In a random sample of 20 students, the mean age is found to be 22.9 years. From past studies, the standard deviation is known to be 1.5 years, and the population is normally distributed. Construct a 90% confidence interval of the pop mean age.



- First find the critical values



$z_c = 1.645$

- Margin of error:

$$E = z_c \frac{\sigma}{\sqrt{n}} = 1.645 \cdot \frac{1.5}{\sqrt{20}} \approx 0.6$$

- Confidence interval:

Left Endpoint:

$$\bar{x} - E$$

$$\approx 22.9 - 0.6$$

$$= 22.3$$

Right Endpoint:

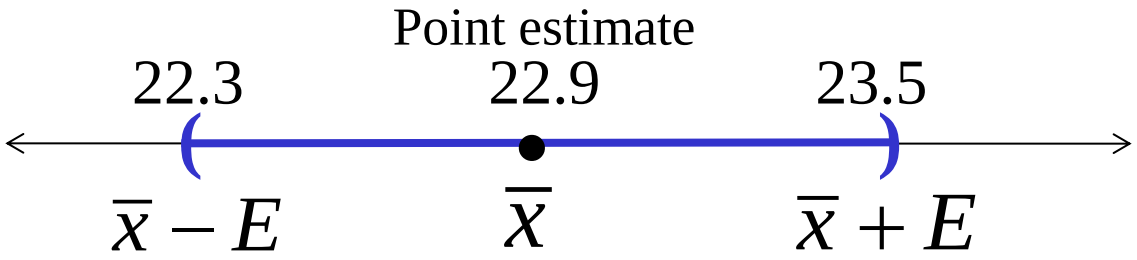
$$\bar{x} + E$$

$$\approx 22.9 + 0.6$$

$$= 23.5$$

$22.3 < \mu < 23.5$

$22.3 < \mu < 23.5$



With 90% confidence, you can say that the mean age of all the students is between 22.3 and 23.5 years.

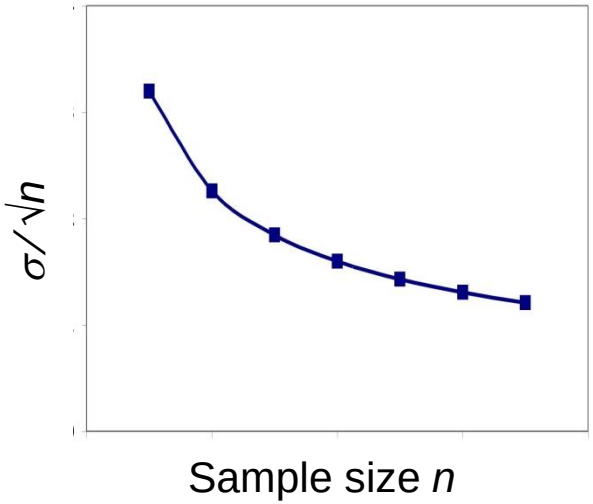
Interpreting the Results

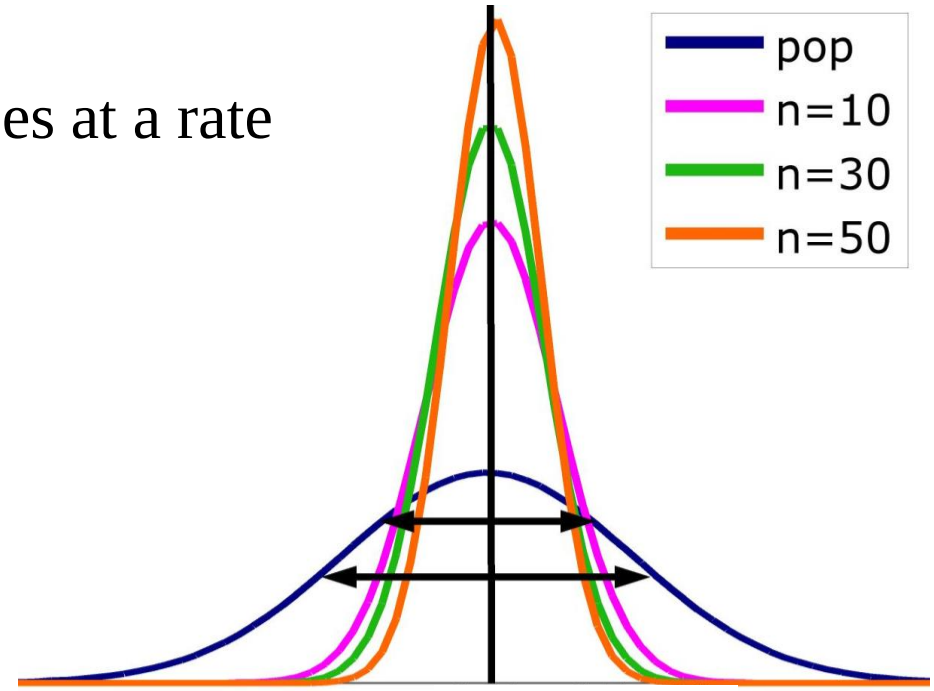
- μ is a fixed number. It is either in the confidence interval or not.
- **Incorrect:** "There is a 90% probability that the actual mean is in the interval (22.3, 23.5)."
- **Correct:** "If a large number of samples is collected and a confidence interval is created for each sample, approximately 90% of these intervals will contain μ.

Impact of sample size

The spread in the sampling distribution of the mean is a function of the number of individuals per sample.

- The larger the sample size, the smaller the standard deviation (spread) of the sample mean distribution.
- But the spread only decreases at a rate equal to $\sqrt{n}$.





Sample Size

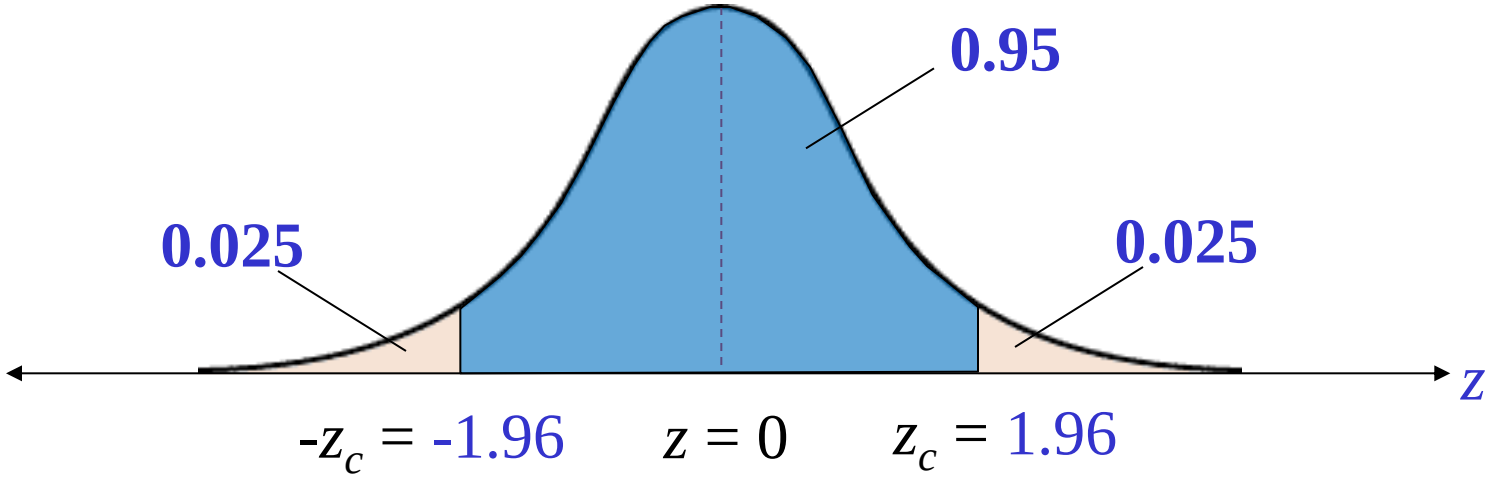
- Given a *c*-confidence level and a margin of error *E*, the minimum sample size *n* needed to estimate the population mean μ is

$$n = \left(\frac{z_c \sigma}{E} \right)^2$$

- If σ is unknown, you can estimate it using *s* provided you have a preliminary sample with at least 30 members.

You want to estimate the mean number of friends for all users of the website. How many users must be included in the sample if you want to be 95% confident that the sample mean is within seven friends of the population mean? Assume the sample standard deviation is about 53.0.

- First find the critical values



$z_c = 1.96$

$z_c = 1.96 \qquad \sigma \approx s = 53.0 \qquad E = 7$

$$n = \left(\frac{z_c \sigma}{E} \right)^2 \approx \left(\frac{1.96 \cdot 53.0}{7} \right)^2 \approx 220.23$$

When necessary, **round up** to obtain a whole number.

You should include **at least 221** users in your sample.

Section 8.2

Confidence Intervals for the Mean (σ Unknown)

The *t*-Distribution

- When the population standard deviation is unknown, the sample size is less than 30, and the random variable x is approximately normally distributed, it follows a ***t*-distribution.**

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

- Critical values of t are denoted by t_c.

Properties of the *t*-Distribution

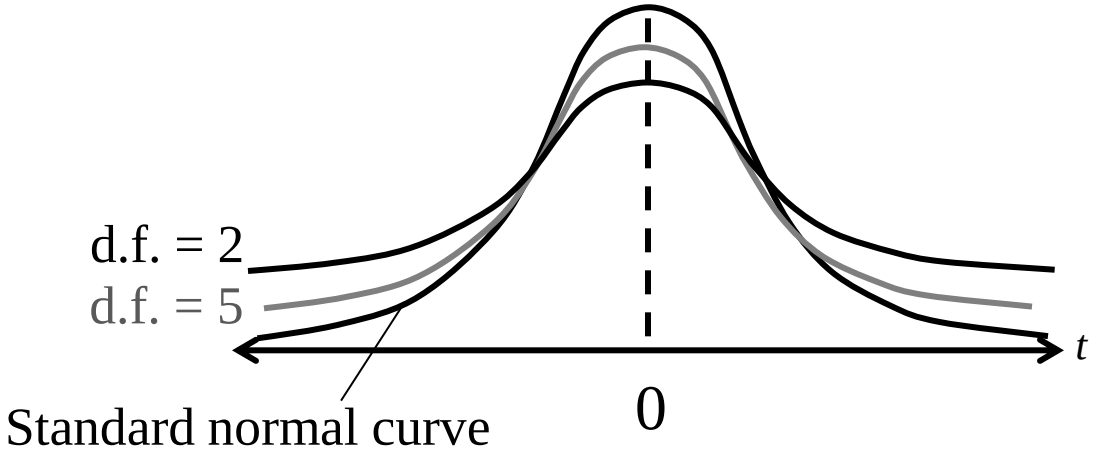
1. The mean, median, and mode of the *t*-distribution are equal to zero.
2. The *t*-distribution is bell shaped and symmetric about the mean.
3. The total area under a *t*-curve is 1 or 100%.
4. The tails in the *t*-distribution are “thicker” than those in the standard normal distribution.
5. The standard deviation of the *t*-distribution varies with the sample size, but it is greater than 1.

Properties of the *t*-Distribution

6. The *t*-distribution is a family of curves, each determined by a parameter called the degrees of freedom. The **degrees of freedom** are the number of free choices left after a sample statistic such as is calculated. When you use a *t*-distribution to estimate a population mean, the degrees of freedom are equal to one less than the sample size.
   - d.f. = $n - 1$ Degrees of freedom

Properties of the *t*-Distribution

7. As the degrees of freedom increase, the *t*-distribution approaches the normal distribution. After 30 d.f., the *t*-distribution is very close to the standard normal *z*-distribution.



Find the critical value t_c for a 95% confidence when the sample size is 15.

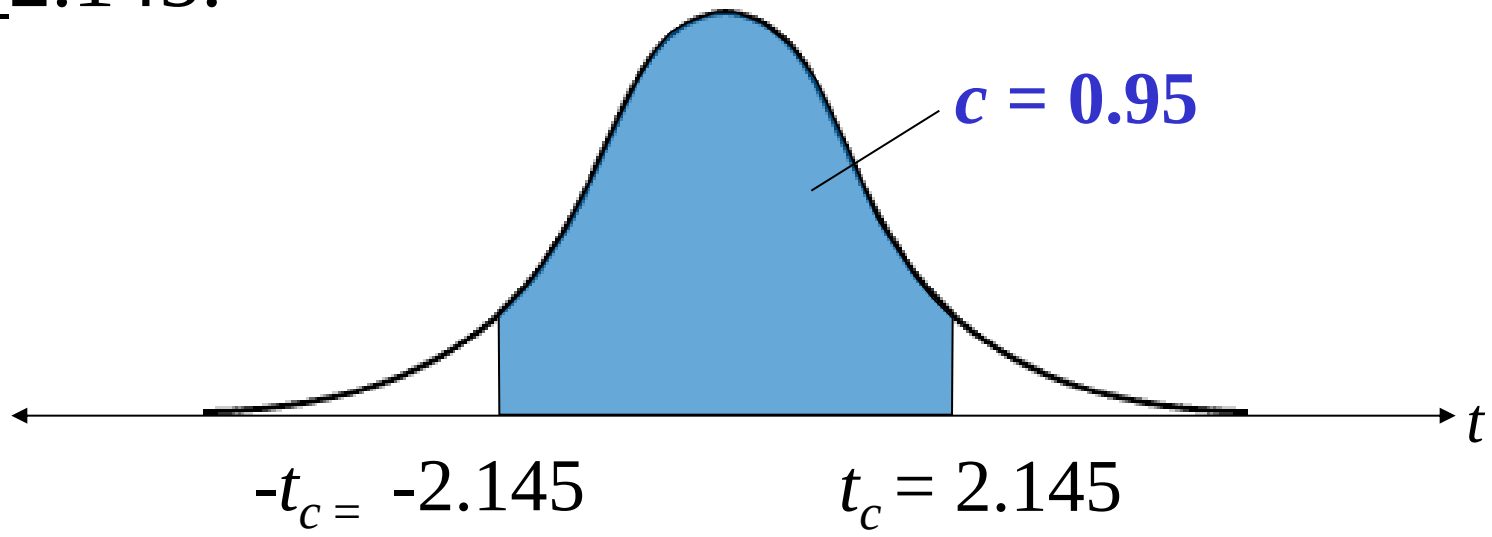
Solution: d.f. = $n - 1 = 15 - 1 = 14$

Table 5: *t*-Distribution

	Level of confidence, c	0.50	0.80	0.90	0.95	0.98	0.99
	One tail, α	0.25	0.10	0.05	0.025	0.01	0.005
d.f.	Two tails, α	0.50	0.20	0.10	0.05	0.02	0.01
1		1.000	3.078	6.314	12.706	31.821	63.657
2		.816	1.886	2.920	4.303	6.965	9.925
3		.765	1.638	2.353	3.182	4.541	5.841
13		.694	1.350	1.771	2.160	2.650	3.012
14		.692	1.345	1.761	2.145	2.624	2.977
15		.691	1.341	1.753	2.131	2.602	2.947
16		.690	1.337	1.746	2.120	2.583	2.921
28		.683	1.313	1.701	2.048	2.467	2.763
29		.683	1.311	1.699	2.045	2.462	2.756
∞		.674	1.282	1.645	1.960	2.326	2.576

$t_c = 2.145$

95% of the area under the *t*-distribution curve with 14 degrees of freedom lies between $t = \pm 2.145$.



Confidence Intervals for the Population Mean

A *c*-confidence interval for the population mean μ

- $\bar{x} - E < \mu < \bar{x} + E \quad$ where $E = t_c \dfrac{s}{\sqrt{n}}$

- The probability that the confidence interval contains μ is c.

You randomly select 16 coffee shops and measure the temperature of the coffee sold at each. The sample mean temperature is 162.0ºF with a sample standard deviation of 10.0ºF. Find the 95% confidence interval for the mean temperature. Assume the temperatures are approximately normally distributed.

Solution:

Use the *t*-distribution ($n < 30$, σ is unknown, temperatures are approximately distributed.)

- n =16, $\bar{x}$ = 162.0 s = 10.0 c = 0.95
- df = n – 1 = 16 – 1 = 15
- Critical Value Table 5: t-Distribution

	Level of confidence, c	0.50	0.80	0.90	0.95	0.98	0.99
	One tail, α	0.25	0.10	0.05	0.025	0.01	0.005
d.f.	Two tails, α	0.50	0.20	0.10	0.05	0.02	0.01
1		1.000	3.078	6.314	12.706	31.821	63.657
2		.816	1.886	2.920	4.303	6.965	9.925
3		.765	1.638	2.353	3.182	4.541	5.841
13		.694	1.350	1.771	2.160	2.650	3.012
14		.692	1.345	1.761	2.145	2.624	2.977
15		.691	1.341	1.753	2.131	2.602	2.947
16		.690	1.337	1.746	2.120	2.583	2.921
28		.683	1.313	1.701	2.048	2.467	2.763
29		.683	1.311	1.699	2.045	2.462	2.756
∞		.674	1.282	1.645	1.960	2.326	2.576

t_c = 2.131

- Margin of error:

$$E = t_c \frac{s}{\sqrt{n}} = 2.131 \cdot \frac{10}{\sqrt{16}} \approx 5.3$$

- Confidence interval:

Left Endpoint:

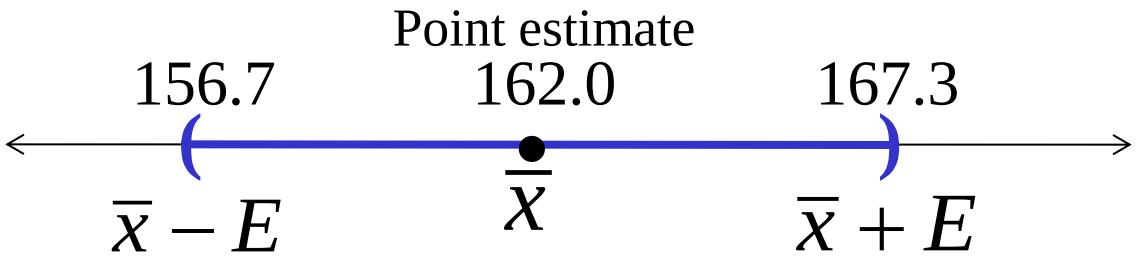
$$\bar{x} - E$$
$$\approx 162 - 5.3$$
$$= 156.7$$

Right Endpoint:

$$\bar{x} + E$$
$$\approx 162 + 5.3$$
$$= 167.3$$

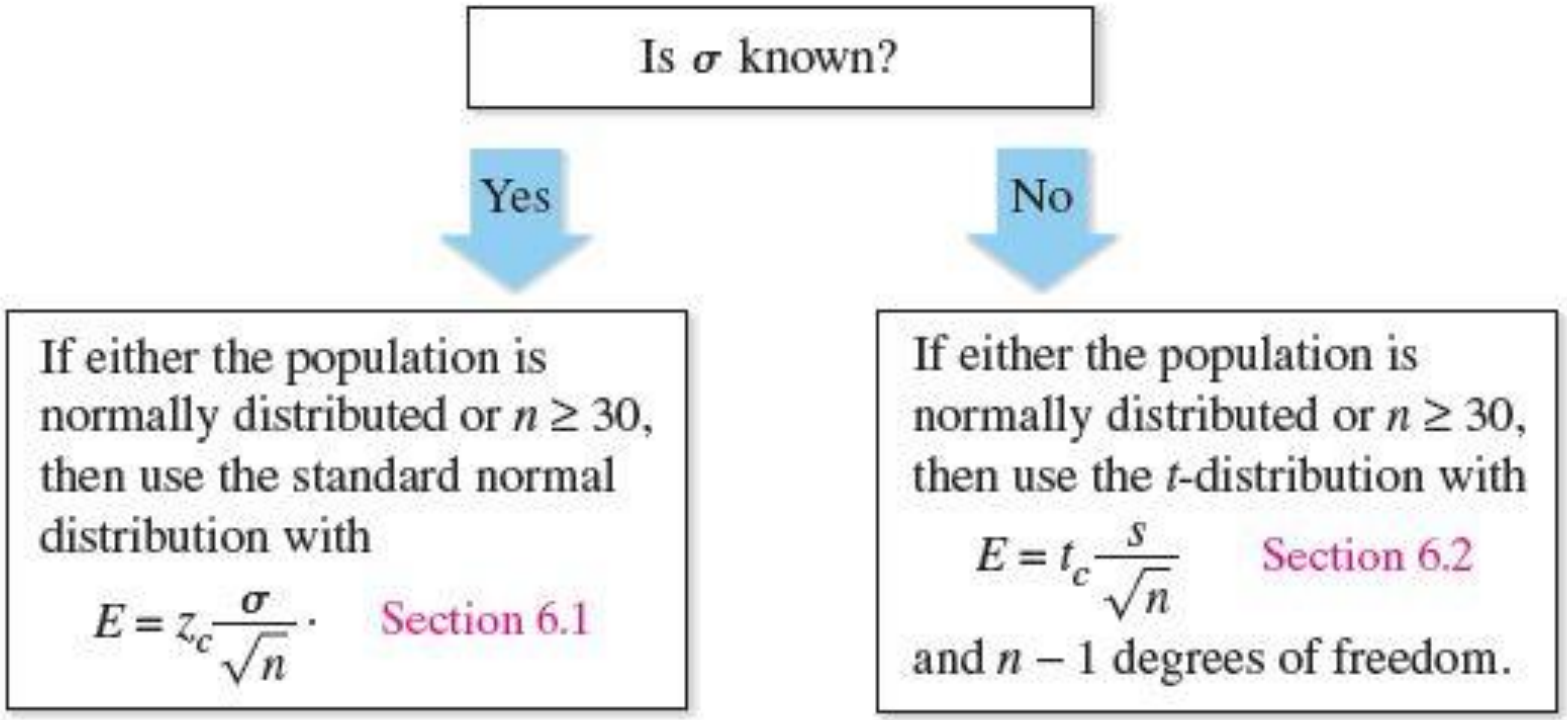
$156.7 < \mu < 167.3$

- **$156.7 < \mu < 167.3$**



With 95% confidence, you can say that the mean temperature of coffee sold is between 156.7ºF and 167.3ºF.

Normal or *t*-Distribution?



You randomly select 25 newly constructed houses. The sample mean construction cost is $181,000 and the population standard deviation is $28,000. Assuming construction costs are normally distributed, should you use the normal distribution, the *t*-distribution, or neither to construct a 95% confidence interval for the population mean construction cost?

Use the normal distribution (the population is normally distributed and the population standard deviation is known)