Lecture 18

The Role of Confidence Intervals in Research

Thought Question 1:

Compare weight loss (over 1 year) in men who diet but do not exercise and vice versa. ***Results***: 95% confidence interval for mean weight loss for men who diet but do not exercise is **13.4 to 18.0 pounds**; 95% confidence interval for mean weight loss for men who exercise but do not diet is **6.4 to 11.2 pounds**.

a. Does this mean 95% of all men who diet will lose between 13.4 and 18.0 pounds? Explain.

b. Do you think you can conclude that men who diet without exercising **lose more weight, on average**, than men who exercise but do not diet?

Thought Question 2:

First confidence interval in Question 1 was based on results from 42 men. Confidence interval spans a range of almost 5 pounds.

If the results had been based on a much ***larger sample***, do you think the **confidence interval** for the mean weight loss would have been **wider, narrower, or about the same**?

Explain your reasoning.

Thought Question 3:

In Question 1, we compared average weight loss for dieting and for exercising by computing separate confidence intervals for the two means and comparing the intervals.

What would be a **more direct value to examine to make the comparison** between the mean weight loss for the two methods?

Thought Question 4:

Case Study 5.3 examined the relationship between baldness and heart attacks. Results expressed in terms of relative risk of heart attack for men with severe vertex baldness compared to men with no hair loss. **95% confidence interval for relative risk** for men under 45 years of age: 1.1 to 8.2.

a. Explain what it means to have a relative risk of 1.1 in this example.

b. Interpret the result given by the confidence interval.

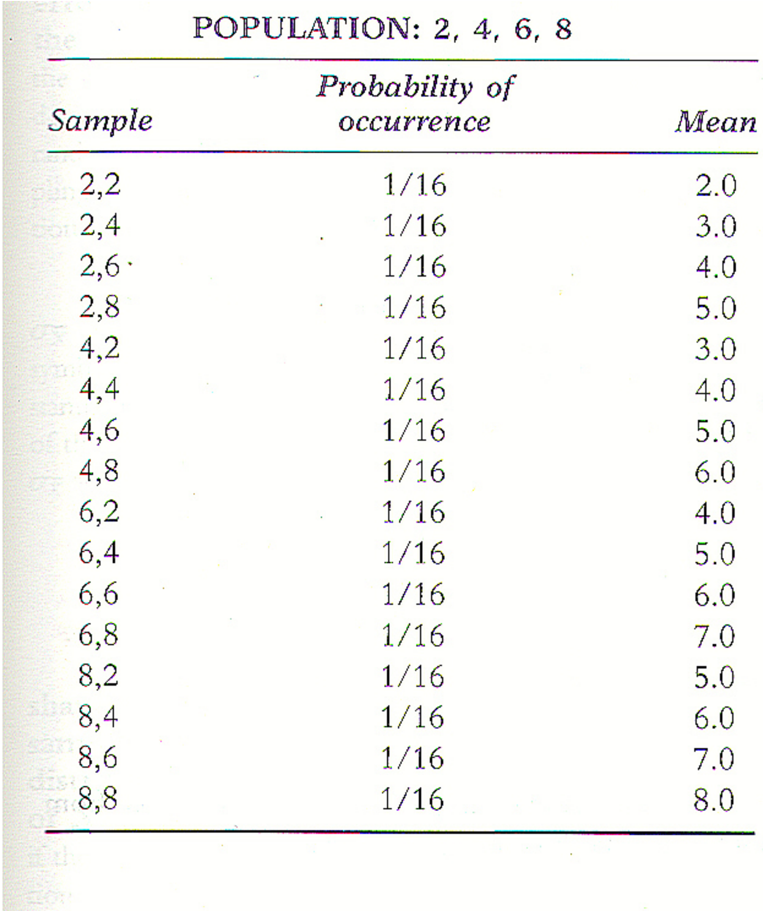
21.1 Confidence Intervals for Population Means

Recall Rule for Sample Means:

If numerous samples or repetitions of same size are taken, the frequency curve of means from various samples will be **approximately bell-shaped**. The **mean** will be same as mean for the population. The **standard deviation** will be:

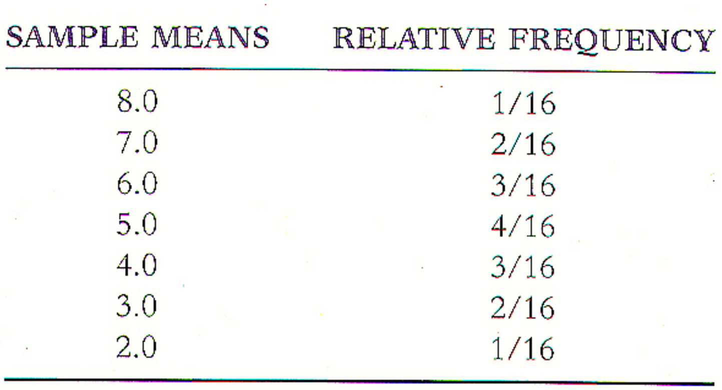
$$\frac{\text{population standard deviation}}{\sqrt{\text{sample size}}}$$

Creating a Sampling Distribution of the Mean



POPULATION: 2, 4, 6, 8

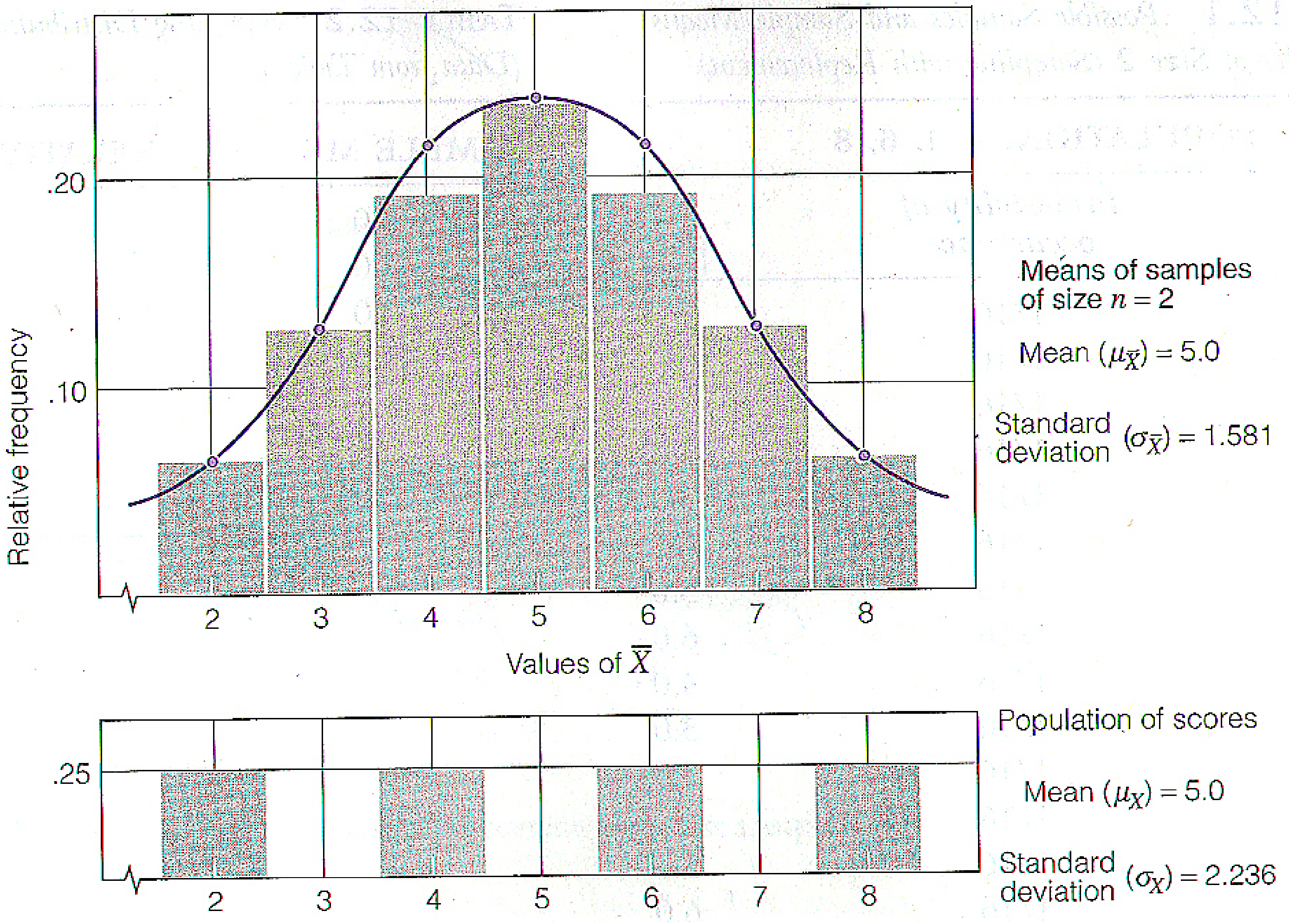
Sample	*Probability of occurrence*	*Mean*
2,2	1/16	2.0
2,4	1/16	3.0
2,6	1/16	4.0
2,8	1/16	5.0
4,2	1/16	3.0
4,4	1/16	4.0
4,6	1/16	5.0
4,8	1/16	6.0
6,2	1/16	4.0
6,4	1/16	5.0
6,6	1/16	6.0
6,8	1/16	7.0
8,2	1/16	5.0
8,4	1/16	6.0
8,6	1/16	7.0
8,8	1/16	8.0



SAMPLE MEANS	RELATIVE FREQUENCY
8.0	1/16
7.0	2/16
6.0	3/16
5.0	4/16
4.0	3/16
3.0	2/16
2.0	1/16

Although there are 16 different possible samples, there are not 16 different sample means possible.

Creating a Sampling Distribution of the Mean



Means of samples of size $n = 2$

Mean $(\mu_{\overline{X}}) = 5.0$

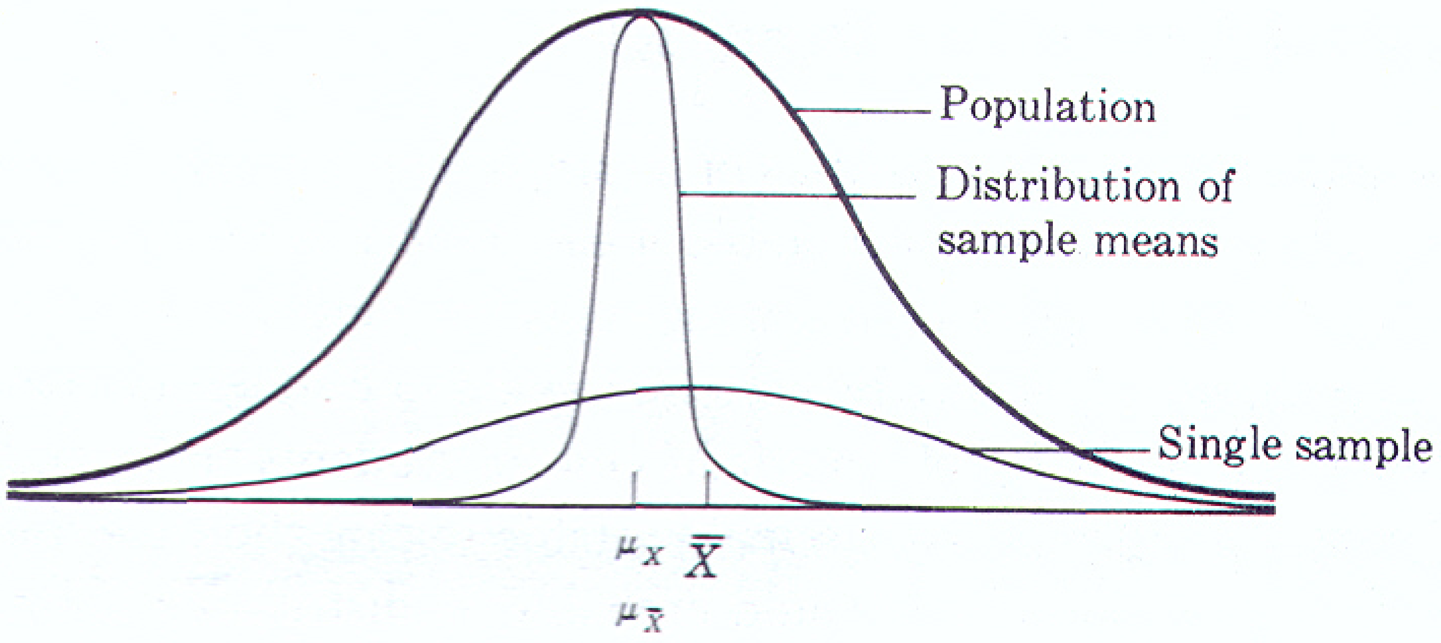
Standard deviation $(\sigma_{\overline{X}}) = 1.581$

Population of scores

Mean $(\mu_X) = 5.0$

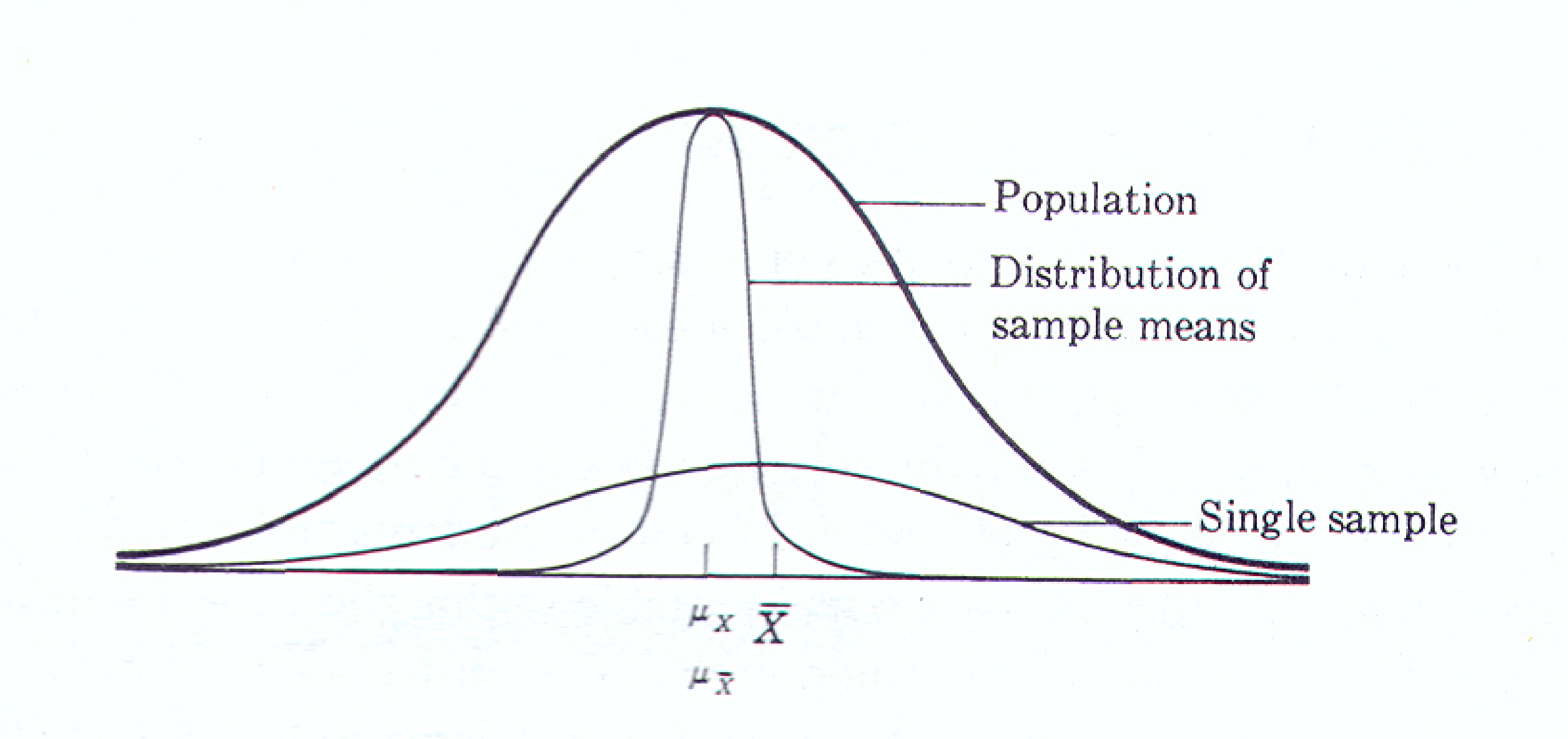
Standard deviation $(\sigma_X) = 2.236$

Sampling Distribution of the Mean



Characteristics of the Sampling Distribution of the Mean

- The mean of the sampling distribution of the mean is about the same as the mean of the original population of individuals.

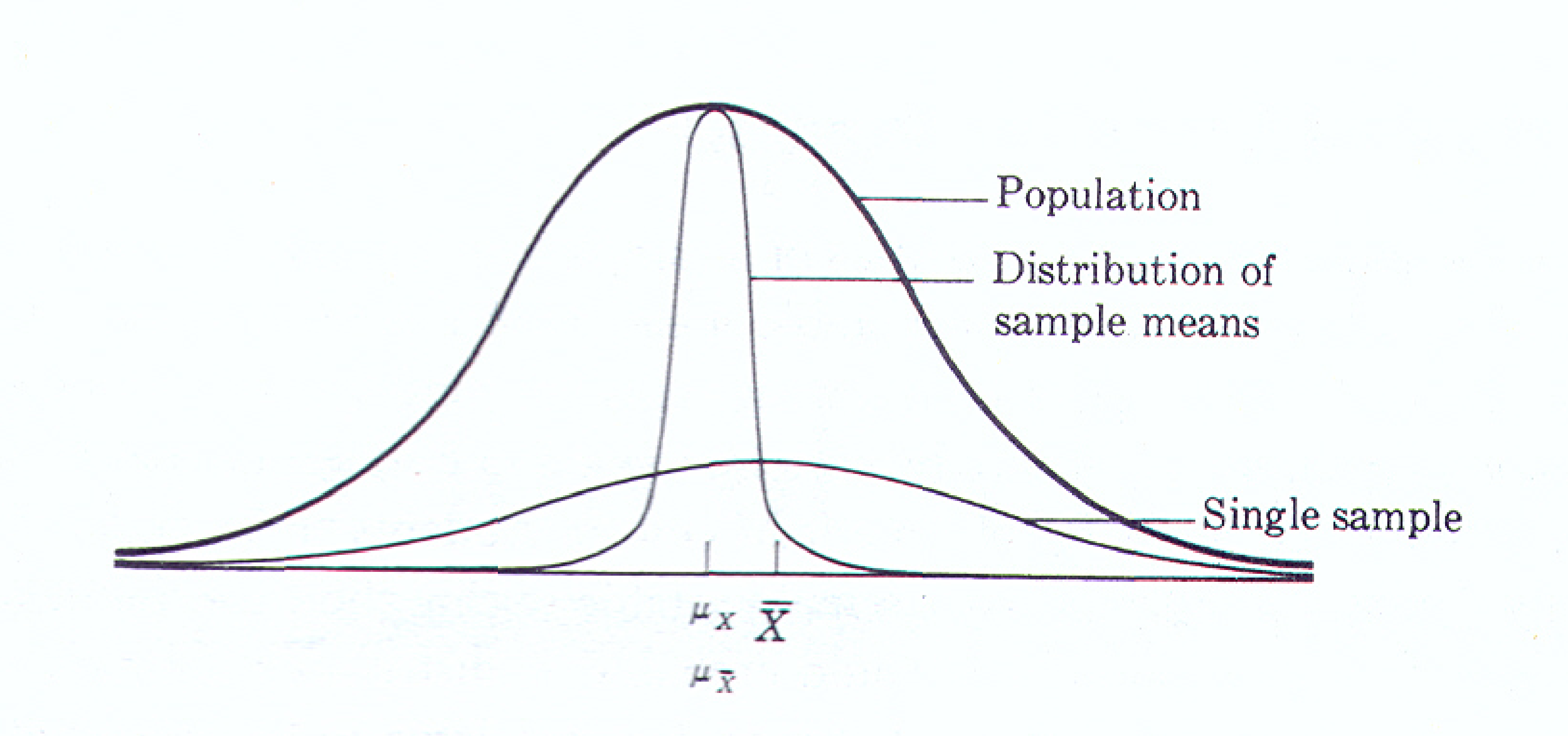


Characteristics of the Sampling Distribution of the Mean

- The mean of the sampling distribution of the mean is about the same as the mean of the original population of individuals.
  - Each sample is based on randomly selected individuals.
  - Thus, the mean of a sample will sometimes be higher and sometimes be lower than the mean of the whole population of individuals.
  - However, because the selection process is random and we are taking a very large number of samples, the high means and low means will, over time, balance each other out.

Characteristics of the Sampling Distribution of the Mean

- The spread of the sampling distribution of the mean is less than the spread of the distribution of the population of individuals.



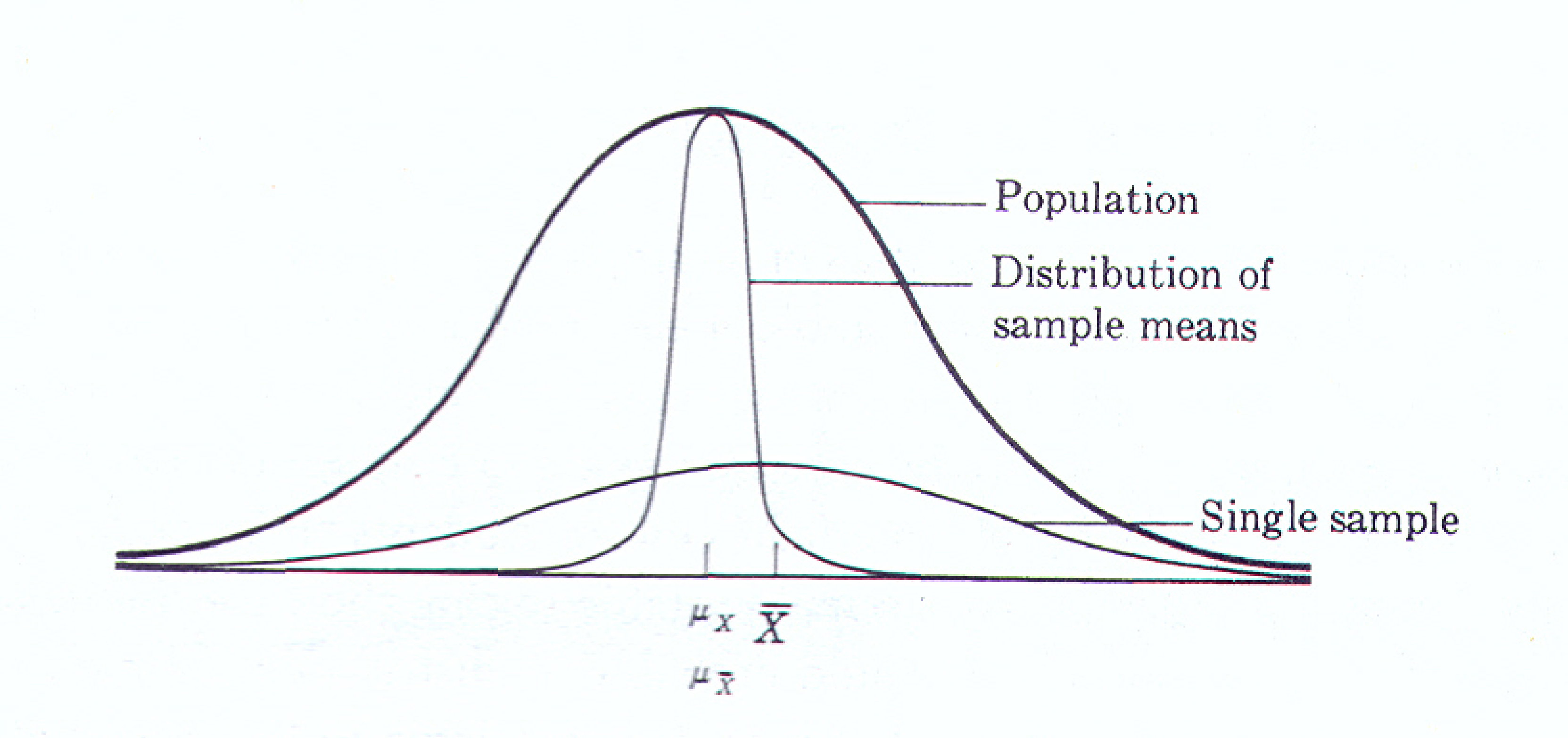
Characteristics of the Sampling Distribution of the Mean

- The spread of the sampling distribution of the mean is less than the spread of the distribution of the population of individuals.
  - Any one score, even an extreme score, has some chance of being included in any random sample.
  - The chance is less of two extreme scores both being included in the same random sample.
  - Further, for a particular random sample to have an extreme mean, the two extreme scores have to be extreme in the same direction (both very high or both very low).

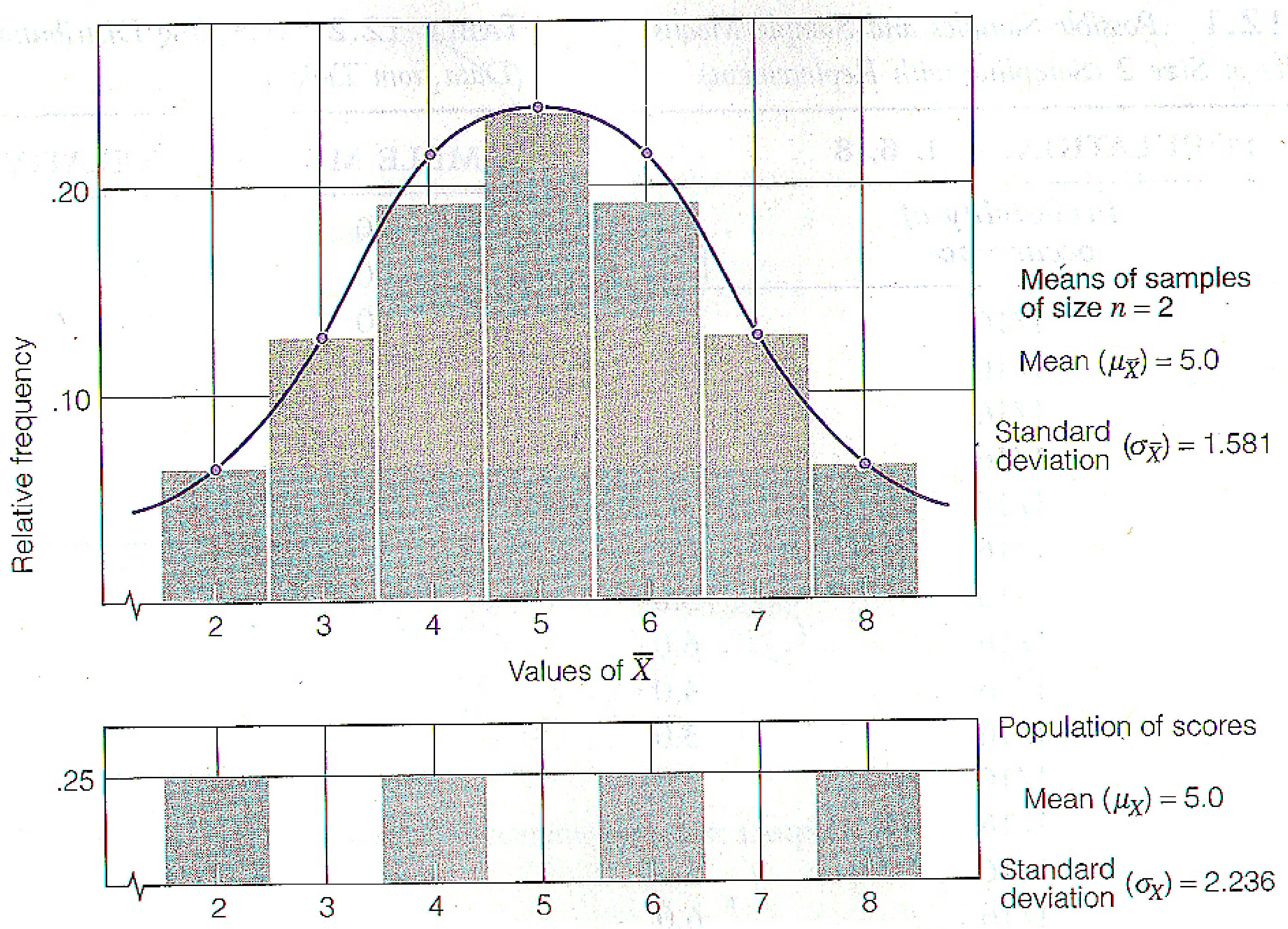
Characteristics of the Sampling Distribution of the Mean

- Thus, having more than a single score in each sample has a moderating effect on the mean of such samples.
- In any one sample, the extremes tend to be balanced out by a middle score or by an extreme in the opposite direction.
- This makes each sample mean tend toward the middle and away from extreme values.
- With fewer extreme means, the variance of the means is less.

Sampling Distribution of the Mean



Sampling Distribution of the Mean



Sampling Distribution of the Mean

- The shape of the distribution of means is approximately normal, if either:
  - Each sample is of 30 or more individuals.
    - Central Limit Theorem
    - Middle scores for means are more likely, and extreme means are less likely with increasing sample size.
  - Or, the distribution of the population of individuals is normal.

Characteristics of the Sampling Distribution of the Mean: Summary

- The mean of the sampling distribution of the mean is about the same as the mean of the original population of individuals.
- The spread of the sampling distribution of the mean is less than the spread of the distribution of the population of individuals.
- The shape of the distribution of means is approximately normal.

Sampling Distribution of the Mean

- In effect, this distribution describes the entire spectrum of sample means that could occur just by chance and thereby provides a frame of reference for generalizing from a single sample mean to a population mean.
- In other words, the sampling distribution of the mean allows us to determine whether, among the set of random possibilities, the one observed sample mean can be viewed as a common outcome or a rare outcome.

Standard Error of the Mean

The standard deviation for the possible sample means is called the **standard error of the mean.** It is sometimes abbreviated by SEM or just "standard error." In other words:

SEM = standard error

= population standard deviation/$\sqrt{n}$

In practice, population standard deviation is unknown and replaced by sample standard deviation, computed from data. Term *standard error of the mean* or *standard error* still used.

Population versus Sample Standard Deviation and Error

Suppose weight losses for thousands of people in a ***population*** were bell-shaped with a mean of 8 pounds and a standard deviation of 5 pounds. A ***sample*** of $n = 25$ people, resulted in a mean of 8.32 pounds and standard deviation of 4.74 pounds.

- population standard deviation = 5 pounds
- sample standard deviation = 4.74 pounds
- standard error of the mean

  (using population S.D.) = $5 / \sqrt{\mathbf{25}} = 1$
- standard error of the mean

  (using sample S.D.) = $4.74 / \sqrt{\mathbf{25}} = 0.95$

Conditions for Rule for Sample Means

1. **Population** of measurements is **bell-shaped**, and a **random sample** of any size is measured.

OR

2. **Population** of measurements of interest is **not bell-shaped**, but a **large random sample** is measured. Sample of size 30 is considered “large,” but if there are extreme outliers, it’s better to have a larger sample.

Constructing a Confidence Interval for a Mean

In 95% *of all samples, the* sample mean *will fall* within 2 standard errors *of the* true population mean.

In 95% *of all samples, the* true population mean *will fall* within 2 standard errors *of the* sample mean.

A 95% confidence interval for a population mean:

sample mean ± 2 standard errors

where standard error = standard deviation/$\sqrt{n}$

Important note: Formula used only if at least 30 observations in the sample. A 95% confidence interval for population mean based on smaller samples requires a multiplier larger than 2, found from a "*t*-distribution."

In order to construct a confidence interval with a different level of confidence, we just need to multiply SEM by another value of z (a Table Value).

Confidence Level	Table Value
90%	1.65
95%	1.96
98%	2.33
99%	2.56

IMPORTANT NOTE:

The formula given above should only be used in there are at least 30 observations in the sample. To compute a confidence interval for the population mean based on smaller samples, a different type of Table Value based on the "t-distribution".

Example #1: A SRS of 200 British men was obtained and the height for each man was recorded. The sample mean was 68.2 inches and the standard deviation was 2.7 inches.

Construct and interpret the 95% confidence interval for the average height of ALL British men.

Example 1: Comparing Diet and Exercise

Compare weight loss (over 1 year) in men who diet but do not exercise and vice versa.

Diet Only Group:

- **sample mean = 7.2 kg**
- **sample standard deviation = 3.7 kg**
- **sample size = *n* = 42**
- **standard error = 3.7/ $\sqrt{42}$ = 0.571**
- **95% confidence interval for population mean:**

$$7.2 \pm 2(0.571) = 7.2 \pm 1.1$$

6.1 kg to 8.3 kg or 13.4 lb to 18.3 lb

Example 1 continued: Exercise Only Group

- **sample mean = 4.0 kg**
- **sample standard deviation = 3.9 kg**
- **sample size = n = 47**
- **standard error = 3.9/ $\sqrt{47}$ = 0.569**
- **95% confidence interval for population mean:**

4.0 ± 2(0.569) = 4.0 ± 1.1

2.9 kg to 5.1 kg or 6.4 lb to 11.2 lb

Appears that dieting results in larger weight loss than exercise because ***no overlap in two intervals***. We are fairly certain average weight loss from dieting is no lower than 13.4 pounds and average weight loss from exercising is no higher than 11.2 pounds.

21.2 Confidence Intervals for Difference Between Two Means

To compare the population means under two conditions or for two groups we could …

1. construct separate confidence intervals for the two conditions and then compare them; or (a better idea)
2. construct a single confidence interval for the *difference* in the population means for the two groups/conditions.

General form for Confidence Intervals:
sample value ± 2 × measure of variability

Constructing a 95% Confidence Interval for the Difference in Means

1. Collect a large sample of observations, independently, under each condition/from each group. **Compute the mean and standard deviation for each sample**.
2. Compute the **standard error of the mean (SEM) for each sample** by dividing the sample standard deviation by the square root of the sample size.
3. Square the two SEMs and add them together. Then take the square root. This will give you the **standard error of the difference in two means.**

   measure of variability $= \sqrt{[(\text{SEM}_1)^2 + (\text{SEM}_2)^2]}$

Constructing a 95% Confidence Interval for the Difference in Means

4. **A 95% confidence interval for the difference in the two population means** is:

   difference in sample means ± 2 × measure of variability

   or

   difference in sample means ± 2 × $\sqrt{[(\text{SEM}_1)^2 + (\text{SEM}_2)^2]}$

Example 2: Comparing Diet and Exercise

Steps 1 and 2. Compute sample means, standard deviations, and SEMs:

Diet Only:

sample mean = 7.2 kg
sample standard deviation = 3.7 kg
sample size = n = 42
standard error = $\mathrm{SEM}_1 = 3.7/\sqrt{42} = 0.571$

Exercise Only:

sample mean = 4.0 kg
sample standard deviation = 3.9 kg
sample size = n = 47
standard error = $\mathrm{SEM}_2 = 3.9/\sqrt{47} = 0.569$

Example 2: Comparing Diet and Exercise

Step 3. Compute standard error of the difference in two means:

$$\text{measure of variability} = \sqrt{[(0.571)^2 + (0.569)^2]} = 0.81$$

Step 4. Compute the interval:

difference in sample means ± 2 × measure of variability

$$[7.2 - 4.0] \pm 2(0.81)$$

$$3.2 \pm 1.6$$

1.6 kg to 4.8 kg ***or*** 3.5 lb to 10.6 lb

Interval is entirely above zero. We can be ***highly confident that there really is a population difference in average weight loss***, with higher weight loss for dieting alone than for exercise alone.

A Caution about Using This Method

This method is **valid only when *independent* measurements are taken from the two groups**.

If matched pairs are used and one treatment is randomly assigned to each half of the pair, the measurements would not be independent. In this case, differences should be taken for each pair of measurements, and then a confidence interval computed for the mean of those differences.

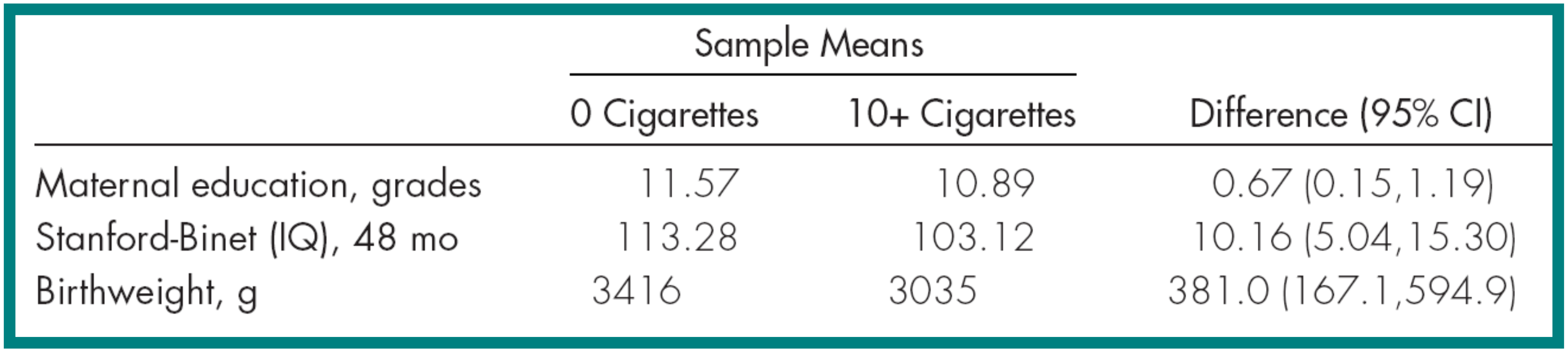
21.3 Revisiting Case Studies: How Journals Present CIs

Direct Reporting of Confidence Intervals: Case Study 6.4

Study of the **relationship between smoking during pregnancy and subsequent IQ of child**.

Journal article (Olds, Henderson, and Tatelbaum, 1994) provided **95% confidence intervals**, most comparing the means for mothers who didn’t smoke and mothers who smoked ten or more cigarettes per day, hereafter called “smokers.”

Case Study 6.4: Direct Reporting of CIs

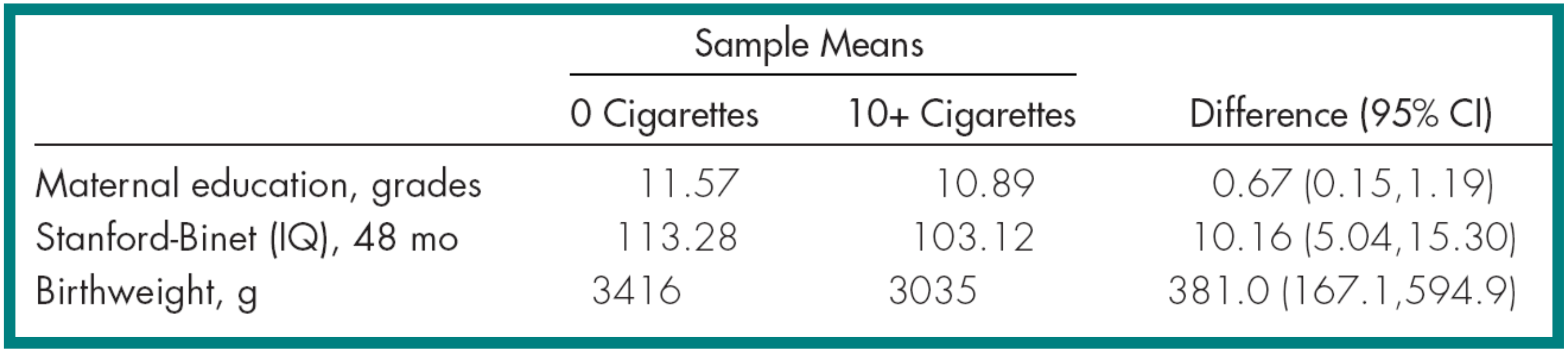


	Sample Means		
	0 Cigarettes	10+ Cigarettes	Difference (95% CI)
Maternal education, grades	11.57	10.89	0.67 (0.15,1.19)
Stanford-Binet (IQ), 48 mo	113.28	103.12	10.16 (5.04,15.30)
Birthweight, g	3416	3035	381.0 (167.1,594.9)

- **Education Interval:** Average educational level for nonsmokers was 0.67 year higher than for smokers, and the difference in the population is probably between 0.15 and 1.19 years of education.

  Mothers who did not smoke also likely to have more education. Maternal education = *confounding variable.*

Case Study 6.4: Direct Reporting of CIs

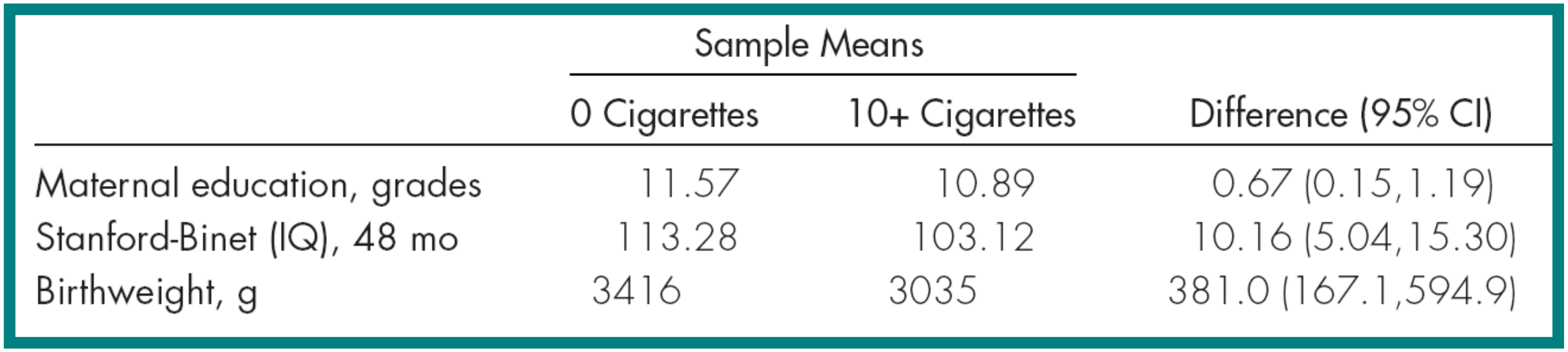


	Sample Means		
	0 Cigarettes	10+ Cigarettes	Difference (95% CI)
Maternal education, grades	11.57	10.89	0.67 (0.15,1.19)
Stanford-Binet (IQ), 48 mo	113.28	103.12	10.16 (5.04,15.30)
Birthweight, g	3416	3035	381.0 (167.1,594.9)

- **IQ Interval:** Difference in means for sample was 10.16 points. There is probably a difference of somewhere between 5.04 and 15.30 points for the entire population.

  Children of nonsmokers in the population probably have IQs that are between 5.04 and 15.30 points higher than the children of mothers who smoke ten or more cigarettes per day.

Case Study 6.4: Direct Reporting of CIs



	Sample Means		
	0 Cigarettes	10+ Cigarettes	Difference (95% CI)
Maternal education, grades	11.57	10.89	0.67 (0.15,1.19)
Stanford-Binet (IQ), 48 mo	113.28	103.12	10.16 (5.04,15.30)
Birthweight, g	3416	3035	381.0 (167.1,594.9)

- **Birthweight Interval:** an explanatory confounding variable; smoking may have caused lower birthweights, which in turn may have caused lower IQs.

  Average difference in birthweight for babies of nonsmokers and smokers in the sample was 381 grams. With 95% confidence, could be a difference as low as 167.1 grams or as high as 594.9 grams for the population.

Case Study 6.4: Direct Reporting of CIs

> “After control for confounding background variables (Table 3), the average difference observed at **12 and 24 months** was 2.59 points (**95% CI: –3.03, 8.20**); the difference observed at **36 and 48 months** was reduced to 4.35 points (**95% CI: 0.02, 8.68**).”
>
> *Source:* Olds and colleagues (1994, pp. 223–224).

From reported confidence intervals:

- Can’t rule out possibility that differences in IQ at 1 and 2 years of age were in other direction because interval covers some negative values.
- Even at 3 and 4 years of age, the CI tells us the gap *could* have been just slightly above zero in the population.

Case Study 6.2:
Reporting Standard Errors of the Mean

Comparison in **serum DHEA-S levels for practitioners and nonpractitioners of transcendental meditation**.

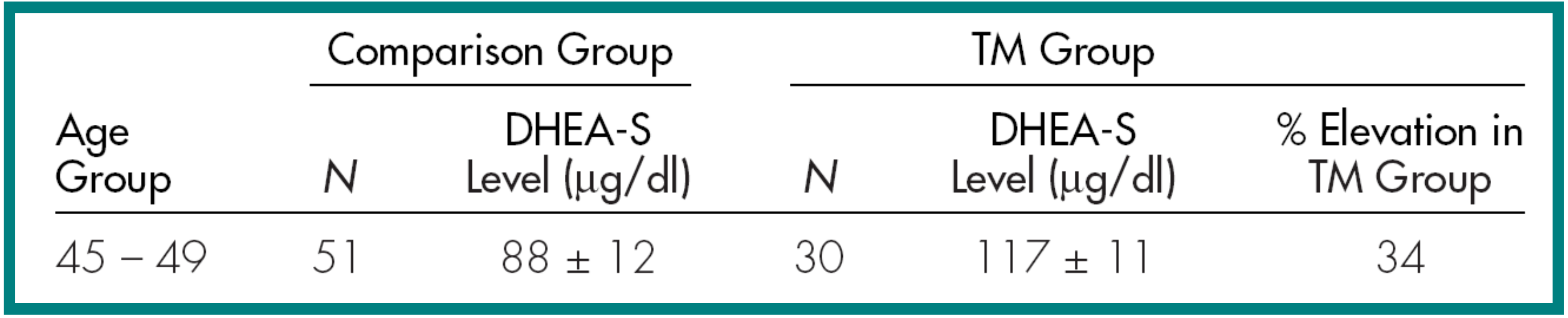
Results presented: mean DHEA-S level for each 5-year age group, separately for men and women.

Confidence intervals not presented, but standard errors of the means (SEMs) were given. So confidence intervals could be computed.

Source: Glaser et al., 1992, p. 333)

Case Study 6.5: Reporting SEMs

Serum DHEA-S Concentrations (± SEM)



	Comparison Group		TM Group		
Age Group	*N*	DHEA-S Level (µg/dl)	*N*	DHEA-S Level (µg/dl)	% Elevation in TM Group
45 – 49	51	88 ± 12	30	117 ± 11	34

$$\text{difference in sample means} \pm 2 \times \sqrt{[(\text{SEM}_1)^2 + (\text{SEM}_2)^2]}$$

$$[7.2 - 4.0] \pm 2 \times \sqrt{[(12)^2 + (11)^2]}$$

$$29 \pm 2(16.3)$$

$$29 \pm 32.6$$

$$-3.6 \text{ to } 61.6$$

Interval includes 0 => cannot say observed difference in sample means represents real difference in population.

Case Study 5.1: Reporting Standard Deviations

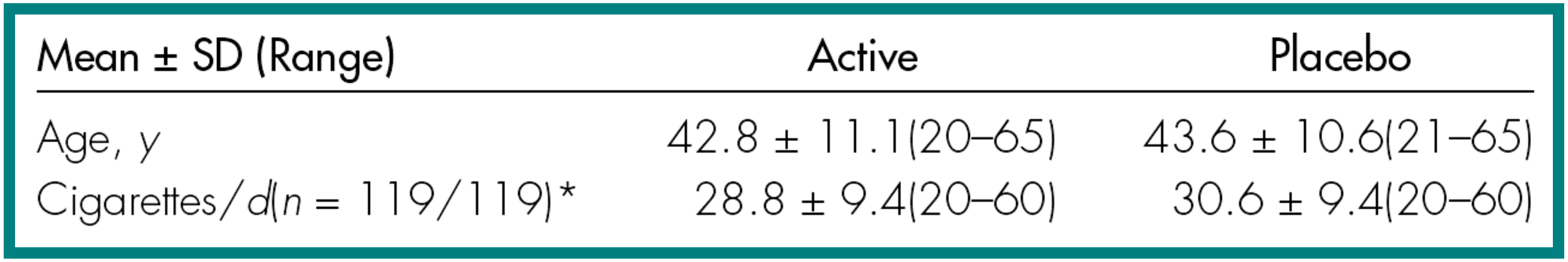
Comparison of **smoking cessation rates** for patients using **nicotine patches** versus **placebo patches**.

Authors reported means, standard deviations (SD), and ranges (low to high) for characteristics to see if the randomization procedure distributed those variables fairly across the two treatment conditions.

Confidence intervals not presented, but could be computed from information provided.

Case Study 5.1: Reporting Std Deviations

Baseline Characteristics



Mean ± SD (Range)	Active	Placebo
Age, *y*	42.8 ± 11.1(20–65)	43.6 ± 10.6(21–65)
Cigarettes/*d*(*n* = 119/119)*	28.8 ± 9.4(20–60)	30.6 ± 9.4(20–60)

*The (*n* = 119/119) => 119 people in each group for these calculations.

Source: Hurt et al., 23 February 1994, p. 596.

Results: slight difference in the mean ages for each group and in the mean number of cigarettes each group smoked per day at the start of the study.

Compute a **95% confidence interval for difference** in mean number of cigarettes smoked per day.

Case Study 5.1: Reporting Std Deviations

Steps 1 and 2. Compute sample means,
standard deviations, and SEMs:

Active Group:

sample mean = 28.8 cigarettes/day
sample standard deviation = 9.4 cigarettes
sample size = n = 119
standard error = $\mathrm{SEM}_1 = 9.4/\sqrt{119} = 0.86$

Placebo Group:

sample mean = 30.6 cigarettes/day
sample standard deviation = 9.4 cigarettes
sample size = n = 119
standard error = $\mathrm{SEM}_2 = 9.4/\sqrt{119} = 0.86$

Case Study 5.1: Reporting Std Deviations

Step 3. Compute standard error of the difference in two means:

$$\text{measure of variability} = \sqrt{[(0.86)^2 + (0.86)^2]} = 1.2$$

Step 4. Compute the interval:

difference in sample means ± 2 × measure of variability

$$[28.8 - 30.6] \pm 2(1.2)$$

$$-1.8 \pm 2.4$$

$$\mathbf{-4.2 \text{ to } 0.60}$$

Could have been slightly fewer cigarettes smoked per day by group that received nicotine patches, but **interval covers zero** => ***can't tell*** if the difference of 1.8 cigarettes observed in the sample means represents a real difference in population means.

Summary of the Variety of Information Given in Journals

Can determine CIs for individual means or difference in two means if you have:

- Direct confidence intervals; or
- Means and standard errors of the means; or
- Means, standard deviations, and sample sizes.

21.4 Understanding Any CI

CI for Relative Risk: Case Study 5.3

Study of the **relationship between baldness and heart disease**. Measure of interest: relative risk of heart disease based on degree of baldness.

> *"For mild or moderate vertex baldness, the age-adjusted RR estimates were approximately 1.3, while for extreme baldness the estimate was 3.4 (95% CI, 1.7 to 7.0). . . . For any vertex baldness (i.e., mild, moderate, and severe combined), the age-adjusted RR was 1.4 (95% CI, 1.2 to 1.9).* *Source*: Lesko et al., 1993, p. 1000.

With 95% certainty men with extreme baldness are estimated to be 1.7 to 7 times more likely to experience a heart attack than men of the same age without any baldness.

Understanding the Confidence Level

For a confidence level of 95%, ***we expect that about 95% of all such intervals will actually cover the true population value***. The remaining 5% will not. Confidence is in the *procedure* over the long run.

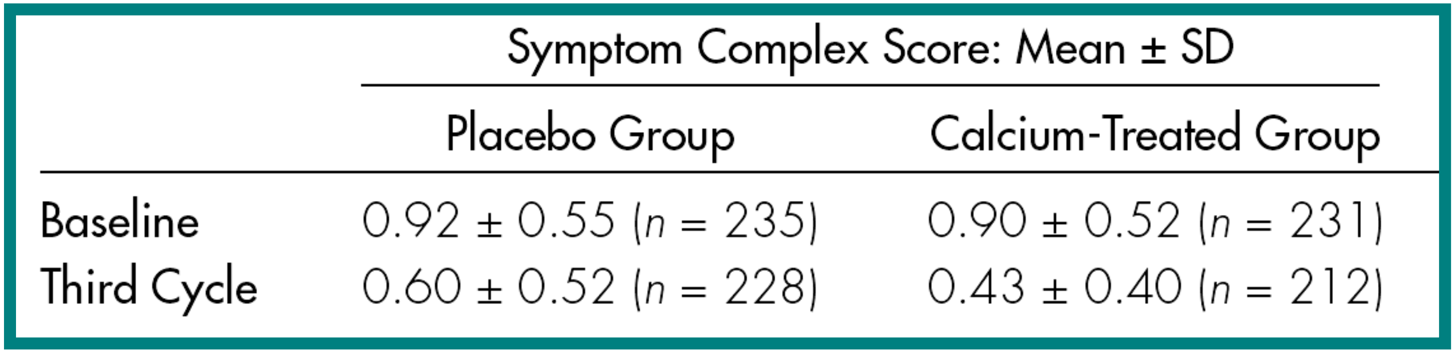
- 90% confidence level => multiplier = 1.645
- 99% confidence level => multiplier = 2.576
- More confidence ⇔ Wider Interval

Case Study 21.1: Premenstrual Syndrome? Try Calcium

- Randomized, double-blind experiment; women who suffered from PMS randomly assigned to either placebo or 1200 mg of calcium per day (4 Tums E-X tablets).
- Participants included 466 women with a history of PMS: 231 in calcium group and 235 in placebo group.
- Response was *symptom complex score* = mean rating (from 0 = absent to 3 = severe) on 17 PMS symptoms.

Source: Thys-Jacobs et al., 1988.

Case Study 21.1: Premenstrual Syndrome? Try Calcium



	Symptom Complex Score: Mean ± SD	
	Placebo Group	Calcium-Treated Group
Baseline	0.92 ± 0.55 (*n* = 235)	0.90 ± 0.52 (*n* = 231)
Third Cycle	0.60 ± 0.52 (*n* = 228)	0.43 ± 0.40 (*n* = 212)

- Difference in means (placebo – calcium) for third cycle is (0.60 – 0.43) = 0.17, and "measure of uncertainty" is 0.039.
- **95% CI for difference** is 0.17 ± 2(0.039), or 0.09 to 0.25.
- Can conclude calcium ***caused*** the reduction in symptoms.
- **Note**: Drop in mean symptom score from baseline to 3rd cycle is about a third for placebo and half for calcium.
- Appears placebos help reduce severity of PMS symptoms too!

For Those Who Like Formulas

Review of Notation from Previous Chapters

Population mean = μ, sample mean = $\overline{X}$, sample standard deviation = s

$z_{\alpha/2}$ = standardized normal score with area α/2 above it

Standard Error of the Mean

Standard error of the mean = SEM = s/ $\sqrt{n}$

Confidence Interval for a Single Population Mean μ

$$\overline{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

Notation for Two Populations and Samples

Population mean = μ_i, i = 1 or 2

Sample mean = $\overline{X}_i$, i = 1 or 2

Sample standard deviation = s_i, i = 1 or 2

Sample size = n_i, i = 1 or 2

Standard error of the mean = $\text{SEM}_i = s_i / \sqrt{n_i}$, i = 1 or 2

Confidence Interval for the Difference in Two Population Means, Independent Samples

$$(\overline{X}_1 - \overline{X}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$