Estimating a Difference in Population Means With Confidence (for Independent Groups)

Mark Rulkowski
Lecturer, Department of Statistics

UNIVERSITY OF MICHIGAN

Research Question

Considering Mexican-American adults (ages 18 - 29) living in the United States, do males and females differ significantly in mean Body Mass Index (BMI)?

- **Population:** Mexican-American adults (ages 18 - 29) in the U.S.
- **Parameter of Interest ($\mu_1 - \mu_2$):** Body Mass Index or BMI (kg/m^2)

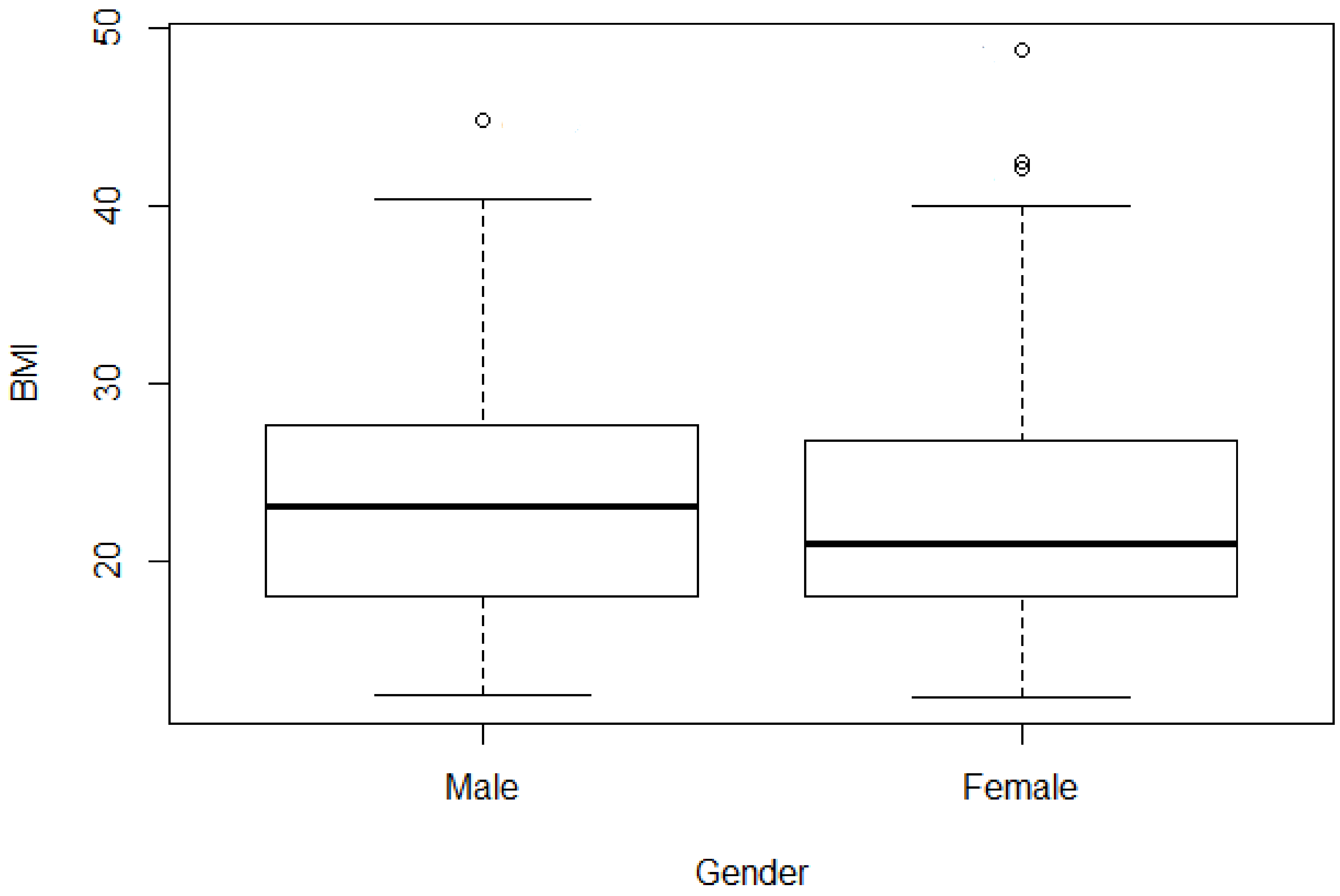
NHANES Data

Gender	BMI	Race	Age 18-29
1	19.9	1	1
2	17.0	1	1
2	26.7	1	1
1	25.6	1	1
…	…	…	…

The data was filtered to include only Mexican-American adults that were between the ages of 18 and 29.

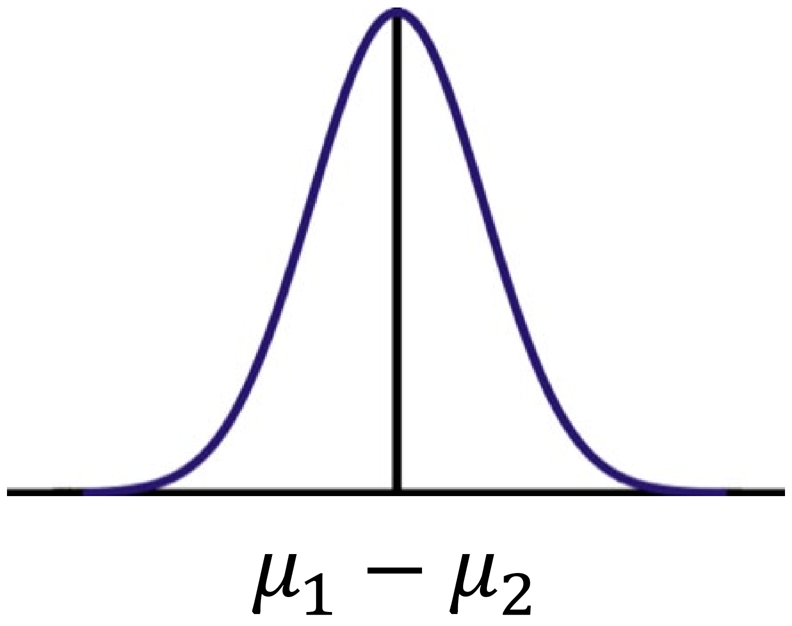
BMI Variable Summary

	Male	Female
Mean	23.57	22.83
St. Dev.	6.24	6.43
Min	12.5	12.4
Max	44.9	48.8
n	258	239



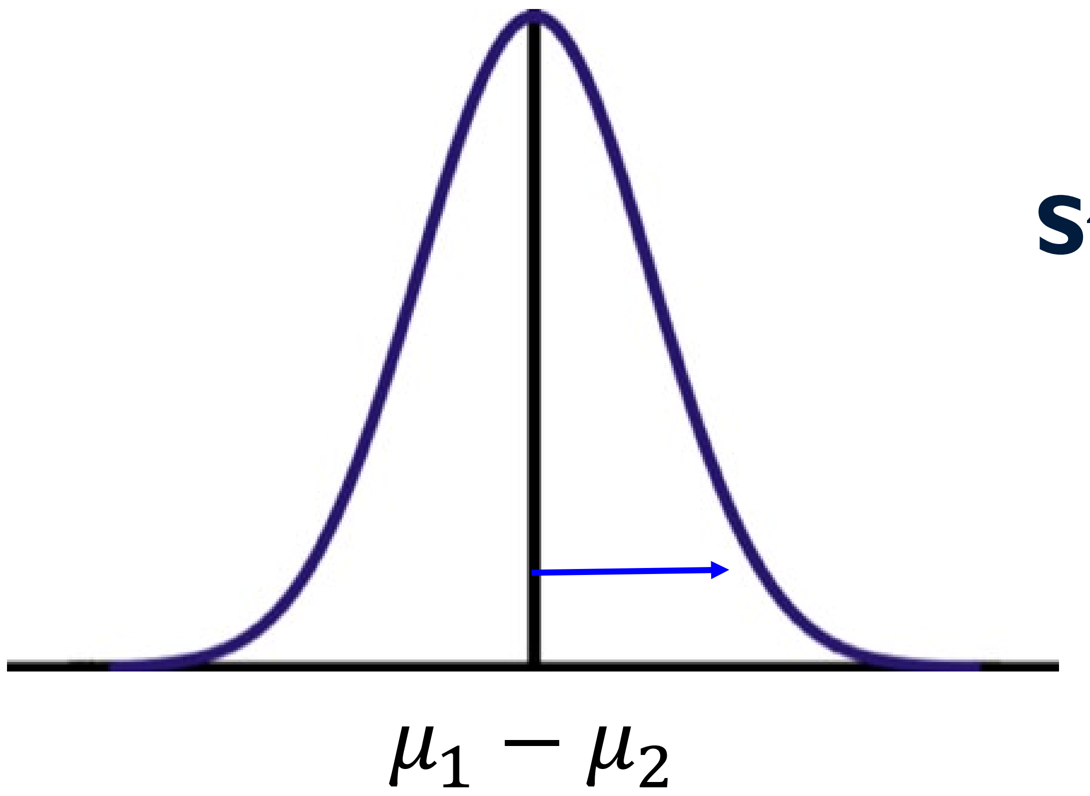
Sampling Distribution of the Difference in Two (Independent) Sample Means

If models for both populations of responses are approximately normal (or sample sizes are both 'large' enough), distribution of the difference in sample means is (approximately) normal.



All possible values of difference in sample means

Sampling Distribution of the Difference in Two (Independent) Sample Means



All possible values of difference in sample means

Standard Error $= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

***Estimated* Standard Error** $= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Confidence Interval Basics

Best Estimate ± Margin of Error

Best Estimate = Unbiased Point Estimate

Margin of Error = "*a few*" Estimated Standard Errors

"a few" = multiplier from appropriate distribution based on desired confidence level and sample design

95% Confidence Level ↔ 0.05 Significance

Confidence Interval Approaches

Pooled Approach

The variance of the two populations are assumed to be equal ($\sigma_1^2 = \sigma_2^2$)

Unpooled Approach

The assumption of equal variances is dropped

Unpooled Confidence Interval Calculations

Best Estimate ± Margin of Error

Difference in sample means ± "a few" · estimated standard error

$$(\bar{x}_1 - \bar{x}_2) \quad \pm \quad t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The df for the t^* multiplier can be found using Welch's approximation.

If technology is not available, a conservative approach can be used by taking the smaller of $n_1 - 1$ and $n_2 - 1$ (i.e. $df = min(n_1 - 1, n_2 - 1)$)

Pooled Confidence Interval Calculations

Best Estimate ± Margin of Error

Difference in sample means ± "a few" · estimated standard error

$$(\bar{x}_1 - \bar{x}_2) \quad \pm \quad \_?\_ \quad \_?\_$$

Pooled Confidence Interval Calculations

Best Estimate ± Margin of Error

Difference in sample means ± "a few" · estimated standard error

$$(\bar{x}_1 - \bar{x}_2) \quad \pm \quad \_?\_ \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Pooled Confidence Interval Calculations

Best Estimate ± Margin of Error

Difference in sample means ± "a few" · estimated standard error

$$(\bar{x}_1 - \bar{x}_2) \quad \pm \quad t^* \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

t^* multiplier comes from a t-distribution with $n_1 + n_2 - 2$ degrees of freedom

Again, this approach can be used if we assume the population variances are equal.

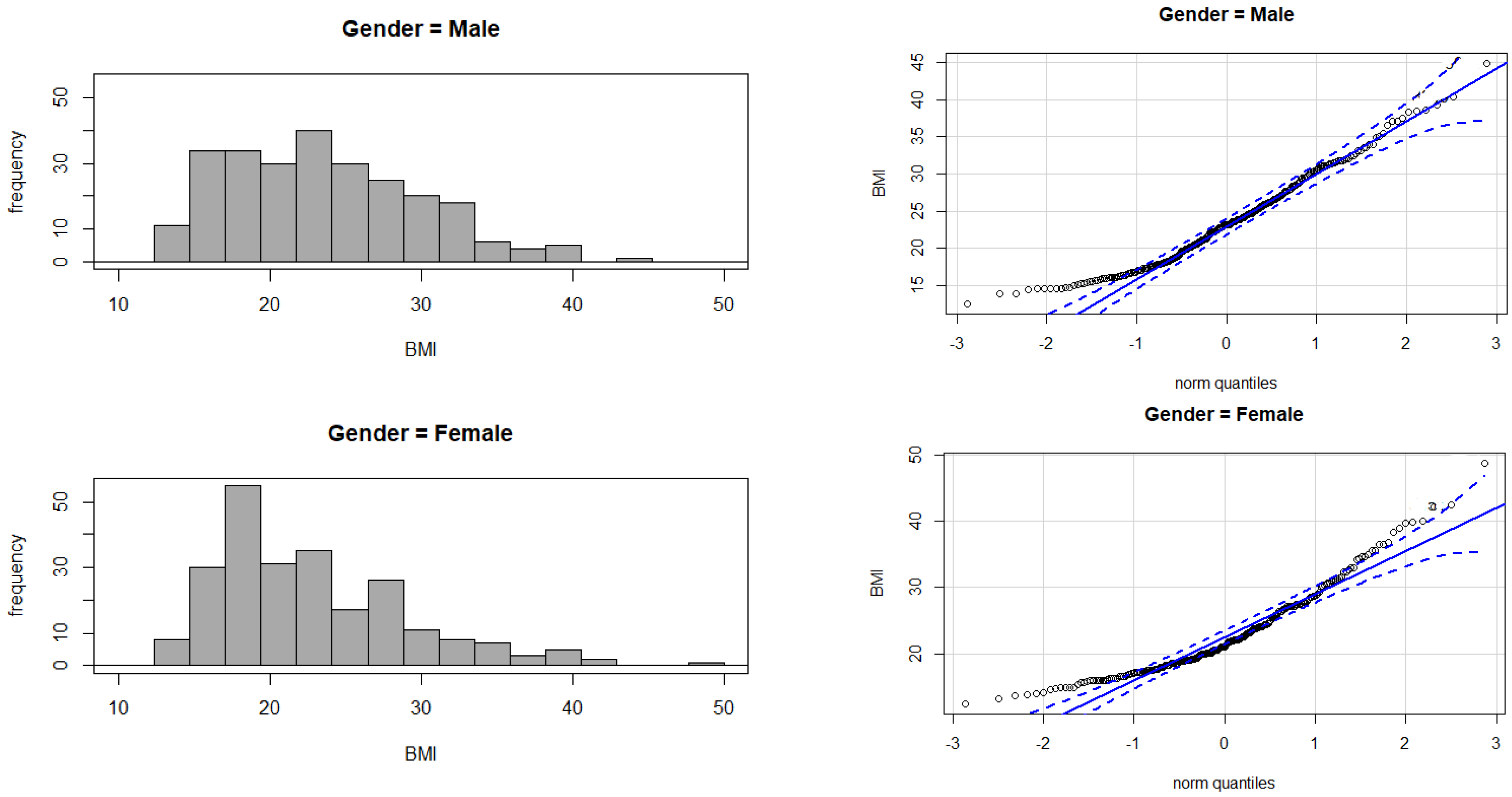
95% Confidence Interval - Example

Considering Mexican-American adults (ages 18 - 29) living in the United States, do males and females differ significantly in mean Body Mass Index (BMI)?

95% Confidence Interval - Example

Normality Assumption: models for both populations of responses are approximately normal (or sample sizes are both 'large' enough)

95% Confidence Interval - Example



95% Confidence Interval - Example

Normality Assumption: models for both populations of responses are approximately normal (or sample sizes are both ‘large’ enough)

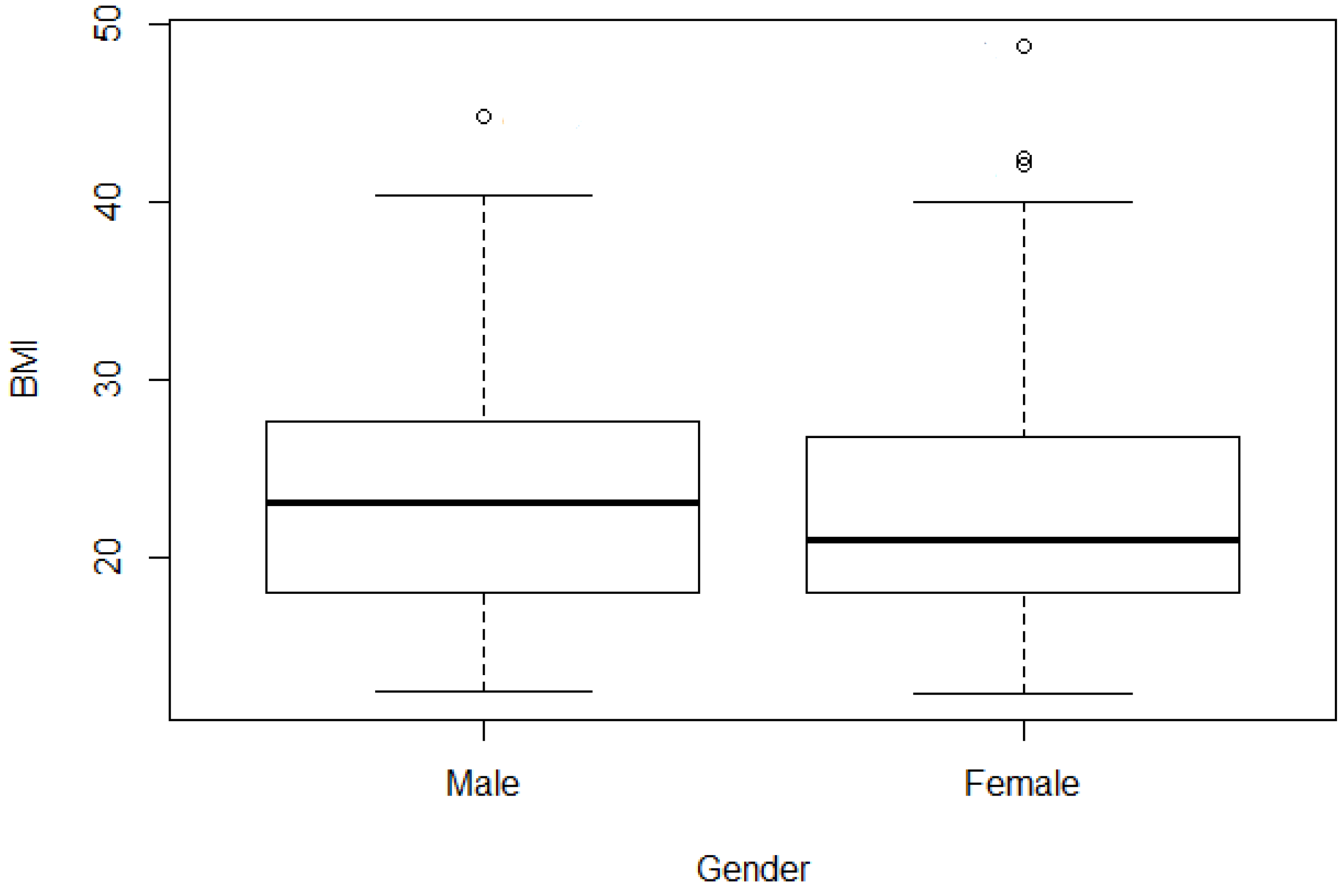
Both distributions have a slight-to-moderate right skew, but the large sample sizes let us apply the CLT and continue.

95% Confidence Interval - Example

Variance Assumption: if we have enough evidence to assume equal variances between the two populations, we can use the "pooled" approach

95% Confidence Interval - Example

	Male	Female
Mean	23.57	22.83
St. Dev.	6.24	6.43
Min	12.5	12.4
Max	44.9	48.8
n	258	239



95% Confidence Interval - Example

Variance Assumption: if we have enough evidence to assume equal variances between the two populations, we can use the "pooled" approach

The IQR's and the standard deviations are similar enough to make this assumptions → the pooled approach will be used!

95% Confidence Interval - Example

	Male	Female
Mean	23.57	22.83
St. Dev.	6.24	6.43
n	258	239

$$(\bar{x}_1 - \bar{x}_2) \quad \pm \quad t^* \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Here, a t^* multiplier of 1.98 will be used

95% Confidence Interval - Example

	Male	Female
Mean	23.57	22.83
St. Dev.	6.24	6.43
n	258	239

$$(23.57 - 22.83) \quad \pm \quad 1.98 \sqrt{\frac{(258-1)6.24^2+(239-1)6.43^2}{258+239-2}} \sqrt{\frac{1}{258} + \frac{1}{239}}$$

$$0.74 \quad \pm \quad 1.98 \quad (6.33) \quad (0.0898)$$

$$0.74 \quad \pm \quad 1.125 \longrightarrow \left(-0.385 \ {}^{kg}/_{m^2}, 1.865 \ {}^{kg}/_{m^2}\right)$$

Interpreting the Confidence **Interval**

$$\left(-0.385\ {}^{kg}/_{m^2}\ , 1.865\ {}^{kg}/_{m^2}\right)$$

"range of reasonable values for our parameter"

With 95% confidence, the difference in mean body mass index between males and females for all Mexican-American adults (ages 18 - 29) in the U.S. is <u>estimated</u> to be between -0.385 kg/m^2 and 1.865 kg/m^2.

Interpreting the Confidence **Level**

What does "with 95% confidence" mean?

If this procedure were repeated over and over,
each time producing a 95% confidence interval estimate,

we would **expect 95% of those resulting intervals**
to contain the difference in population mean BMI.

Summary

- Confidence Intervals are used to give an *interval* estimate for our parameter of interest **~ difference in population means**
- Center of the Confidence Interval is our best estimate **~ difference in sample means**
- Margin of Error is "a few" (estimated) standard errors **~ for two means we use t* multipliers (pooled vs. unpooled)**
- Assumptions for CI's for Difference in Population Means
  ~ data are two simple random samples, independent
  ~ both populations of responses are normal (else n large helps)
- Know how to interpret the **interval** and the **level**