MODULE 10

ANALYSIS OF VARIANCE (ANOVA)

OUTLINE

1. ANOVA overview
2. Significance testing steps
3. Statistical assumptions
4. Pairwise (two-group) comparisons
5. Study questions

OUTLINE

1. ANOVA overview
2. Significance testing steps
3. Statistical assumptions
4. Pairwise (two-group) comparisons
5. Study questions

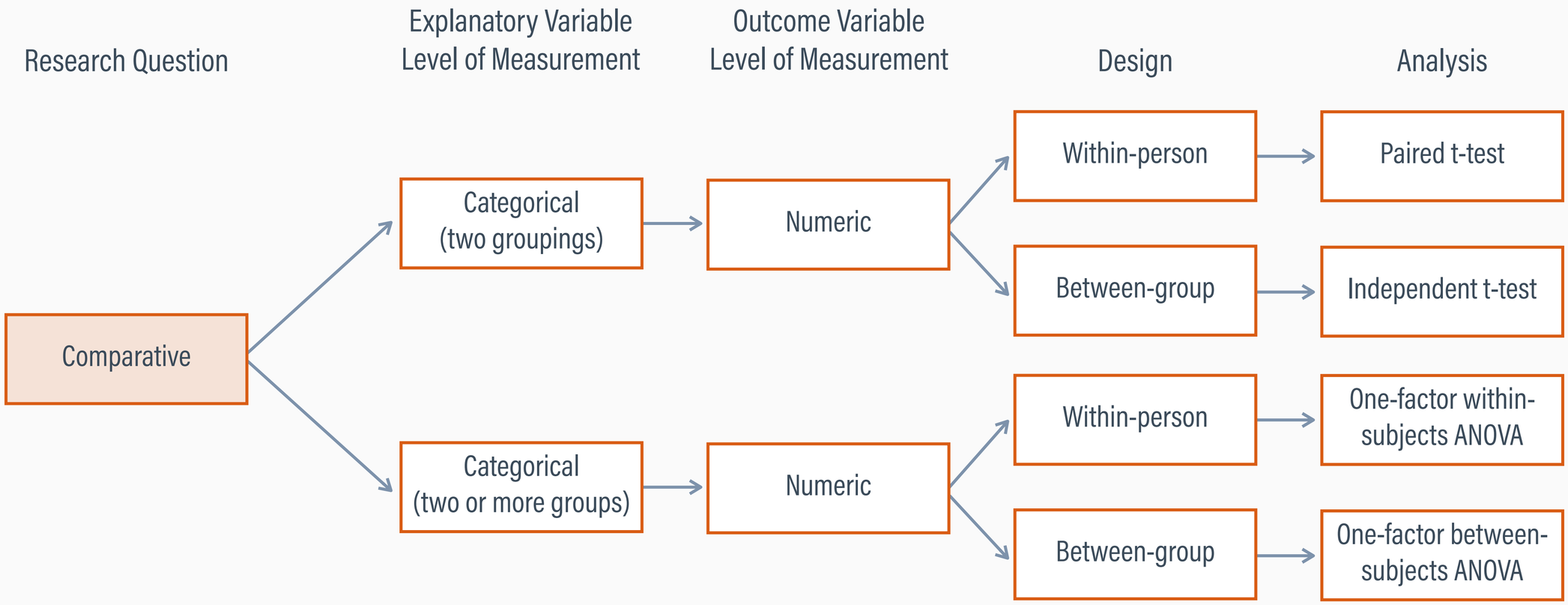
ANALYSIS OF VARIANCE (ANOVA)

- Paired and independent-samples t-tests were limited to comparisons involving two groups or two measurements
- ANOVA is more flexible because it allows for comparisons involving two or more means
- Applicable to comparative research questions and hypotheses involving the difference between two or more means obtained from the different people (between-group designs) or the same people (within-person designs)

ANOVA, CONTINUED

- Separates (partitions) variation into score differences due to group membership and residual (leftover) score differences
- An F-statistic (similar to a t-statistic in logic) compares the size of the group differences (the “effect”) to leftover variation
- ANOVA evaluates whether observed mean differences are larger than expected by chance

STATISTICAL ORG CHART



OUTLINE

1. ANOVA overview
2. Significance testing steps
3. Statistical assumptions
4. Pairwise (two-group) comparisons
5. Study questions

RACIAL DISPARITIES IN PTSD

People of color in the United States disproportionately bear the burden of trauma and posttraumatic stressdisorder (PTSD). Pregnant women of color are at particular risk, as perinatal PTSD is associated with adverse maternal and child health. However, PTSD is a heterogeneous disorder comprising discrete symptom dimensions. In a large sample of Latina, Black, and non-Hispanic White postpartum women in the United States (N = 1663), we examined racial and ethnic differences in the factors of the dysphoric arousal model—a leading dimensional model of PTSD.

Jennifer Thomas

Jennifer Sumner

Thomas, J.L., Carter, S.E., Dunkel-Schetter, Sumner, J.A. (2021). Racial and ethnic disparities in posttraumatic psychopathology among postpartum women. *Journal of Psychiatric Research*, *137*, 36-40.

KEY VARIABLES

Racial Ethnic Identity

The study classified woman into three racial groups: Black, Latina, and non-Hispanic White

Anxious Arousal (Hyperarousal)

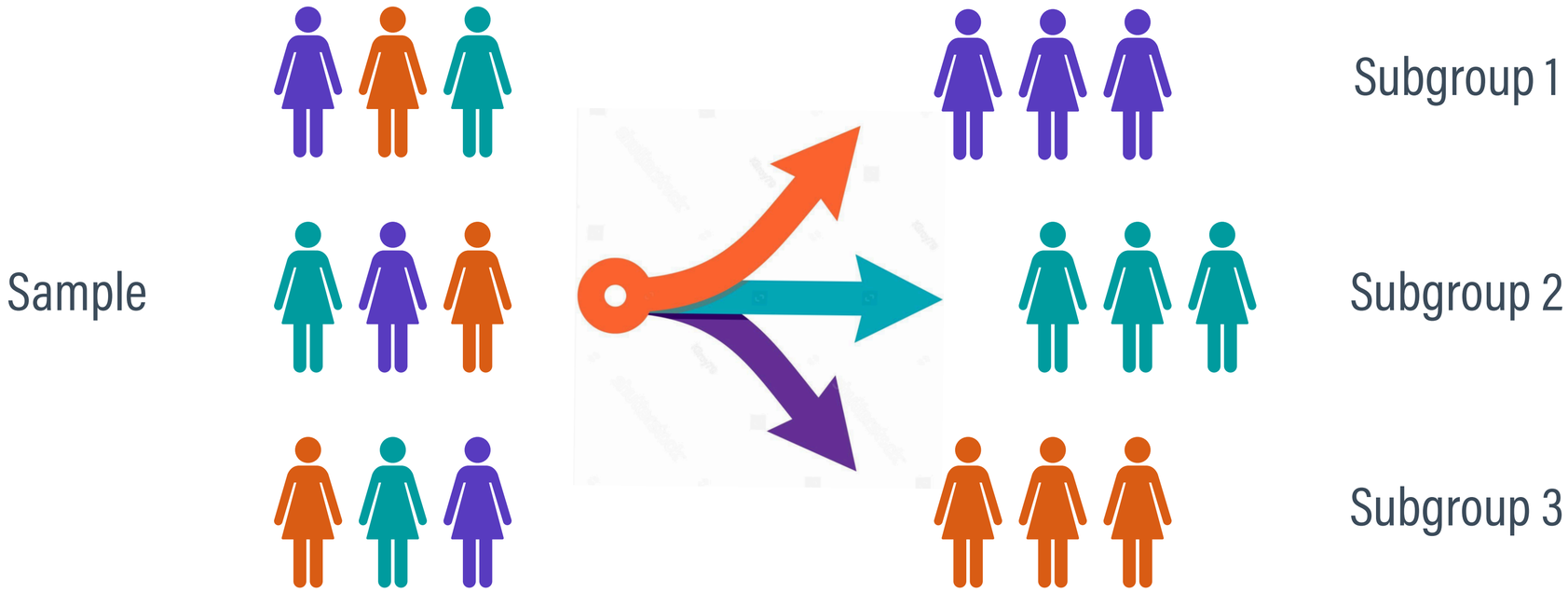
Anxious arousal (hyperarousal) is the abnormally heightened state of anxiety that occurs when thinking about a traumatic event (e.g., hypervigilance, exaggerated startle).

RESEARCH QUESTION

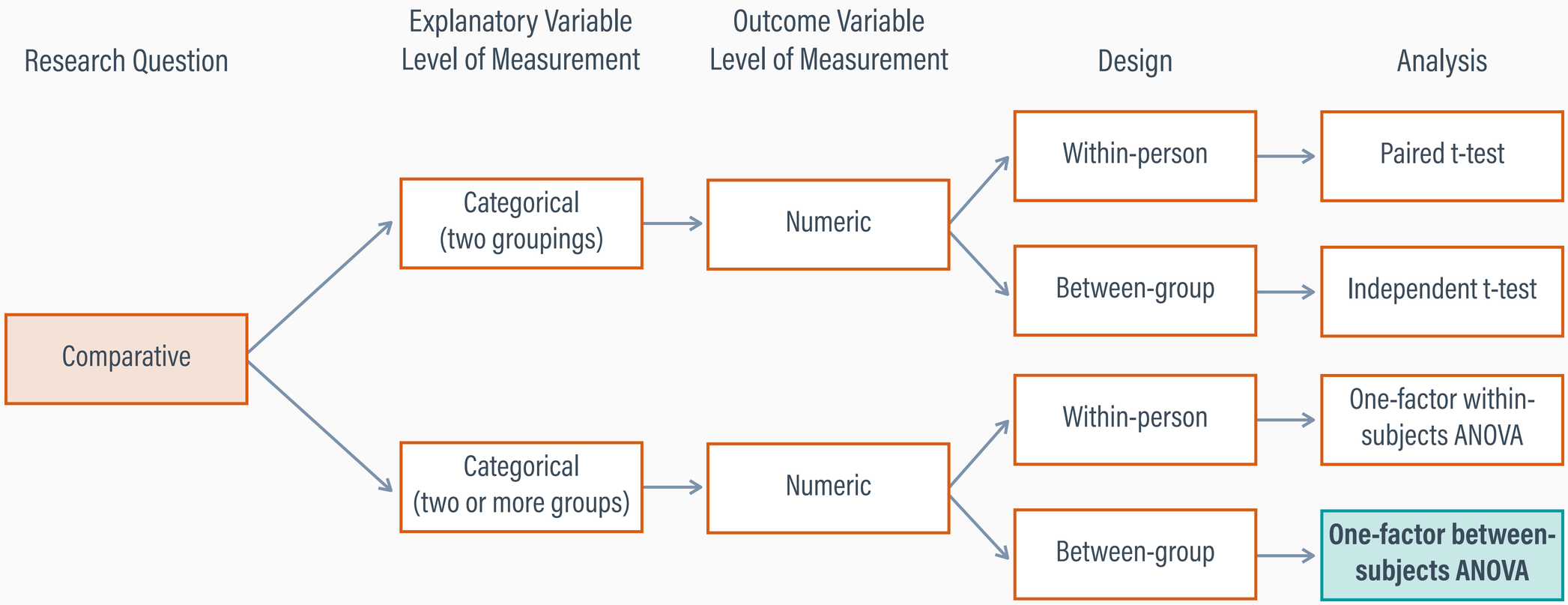
- Question: Do women of color (Black and Latina) differ from White women in their levels of anxious arousal?
- The explanatory (independent) variable, race, consists of three groups: Black, Latina, and White
- The outcome (dependent) variable, anxious arousal (PTSD), is a numeric scale derived from several questionnaire items

BETWEEN-GROUP DESIGN

- Participants divide into subgroups based on a shared qualitative characteristic (independent variable, feature)



STATISTICAL ORG CHART



SIGNIFICANCE TESTING STEPS

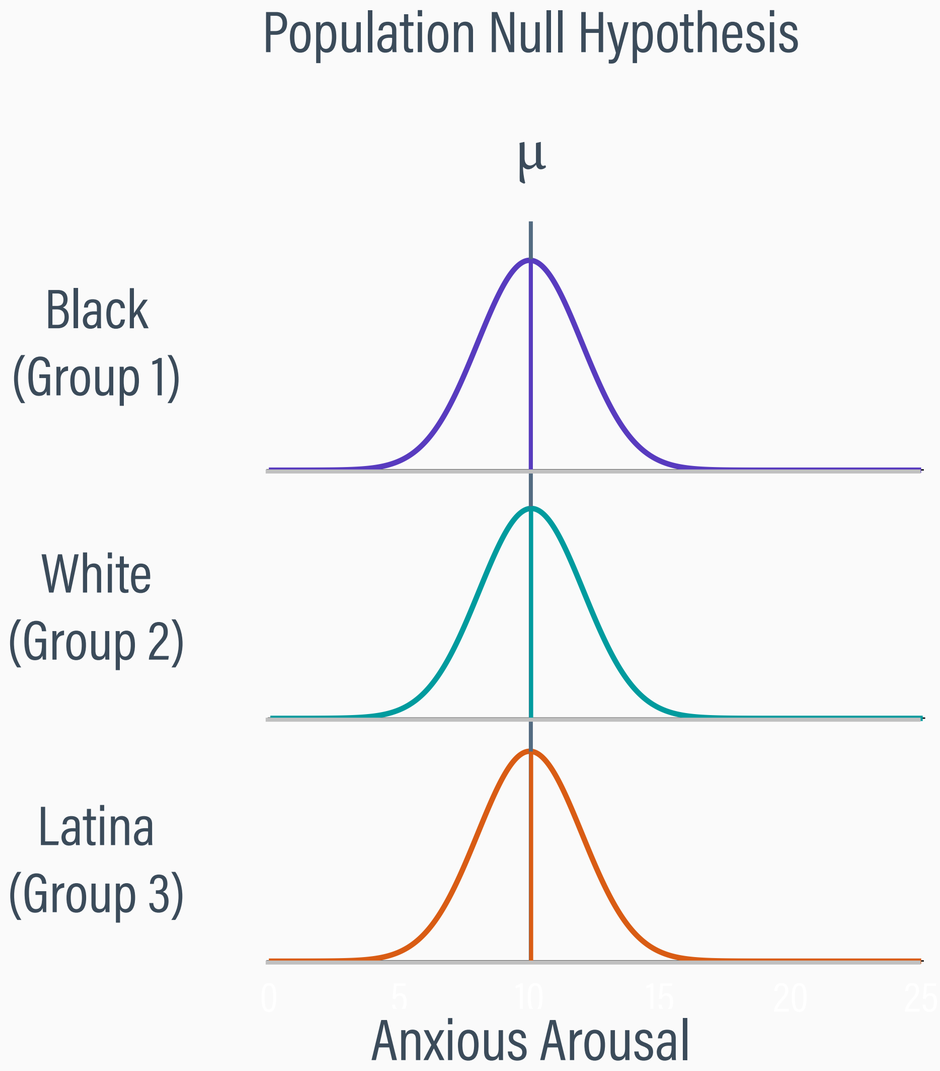
1. Specify hypotheses
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

NULL HYPOTHESIS

- Null hypothesis: In the population, PTSD levels are identical across racial groups

$$H_0: \mu_1 = \mu_2 = \mu_3$$

- The null is counter to expectations because researchers anticipate that racial groups could differ in PTSD



ALTERNATE HYPOTHESIS

- With three groups, the concept of a one-tailed or two-tailed alternate is no longer relevant because there are numerous ways groups could differ

- Alternate hypothesis: In the population, the PTSD means differ across racial groups

$$H_A: \mu_1 \neq \mu_2 \quad \text{and/or} \quad \mu_1 \neq \mu_3 \quad \text{and/or} \quad \mu_2 \neq \mu_3$$

SIGNIFICANCE TESTING STEPS

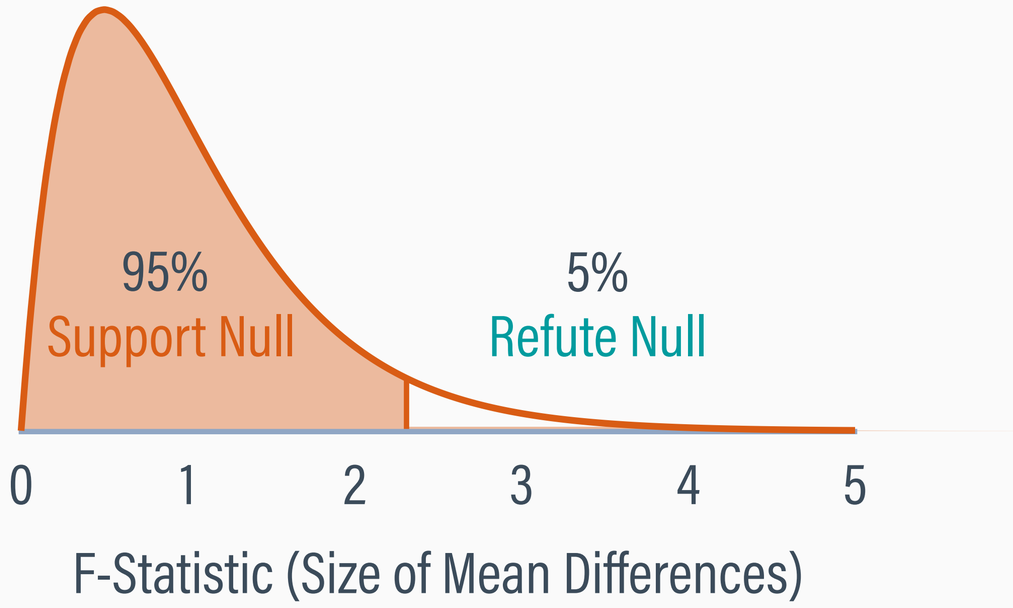
1. Specify hypotheses
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

STANDARD OF EVIDENCE

- The data are the evidence that we use to conclude whether the null is plausible (“innocent”) or implausible (“guilty”)
- If the sample mean differences from our data are very different from zero, we conclude that the null hypothesis is implausible
- ANOVA applies the same logic, but it uses an F-statistic to package all mean differences into a single quantity

5% REJECTION REGION (ALPHA LEVEL)

- By convention, we refute the null if the sample mean differences (encoded by an F-statistic) fall in the most extreme 5% of the sampling distribution
- Such as sample has less than a 5% chance of originating from the null population ($p < .05$)
- The 5% rejection region (alpha level) is only in the upper tail because the F-statistic is always positive—larger values indicate stronger evidence against H_0

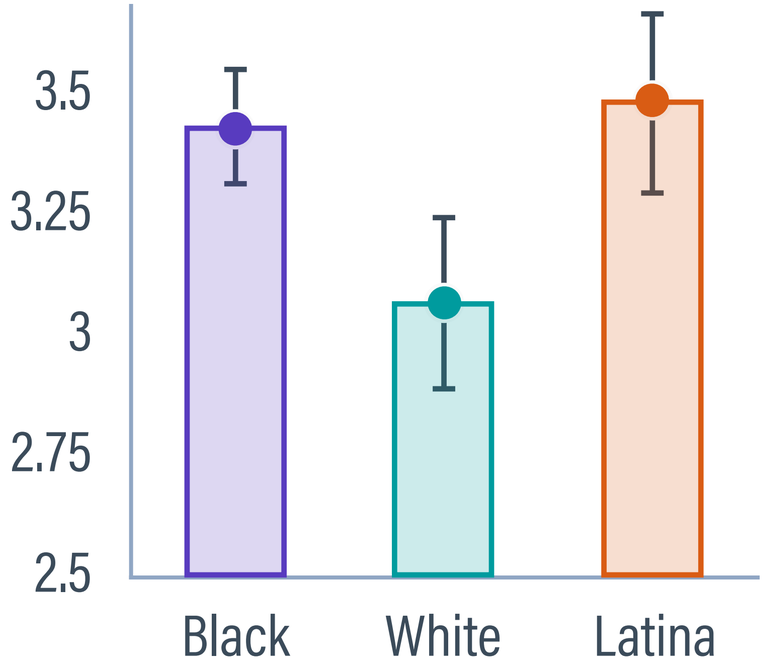


SIGNIFICANCE TESTING STEPS

1. Specify hypotheses
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

Group	$\bar{X}$	s	n
Black	3.43	1.96	901
White	3.07	1.56	389
Latina	3.49	1.61	373

Examine the descriptive statistics from the study. In small groups of two or three, discuss the degree to which the results appear consistent or inconsistent with the null hypothesis. Why?



R OUTPUT

```
 Descriptive statistics by group
Race: Black
                vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
AnxiousArousal     1 901 3.43 1.96      3    3.19 1.48   1  11    10 1.14     0.88 0.07
------------------------------------------------------------------------------------
Race: Latina
                vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
AnxiousArousal     1 373 3.49 1.61      3    3.32 1.48   1  10     9 0.88     0.47 0.08
------------------------------------------------------------------------------------
Race: White
                vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
AnxiousArousal     1 389 3.07 1.56      3    2.85 1.48   1  10     9 1.44     2.56 0.08
```

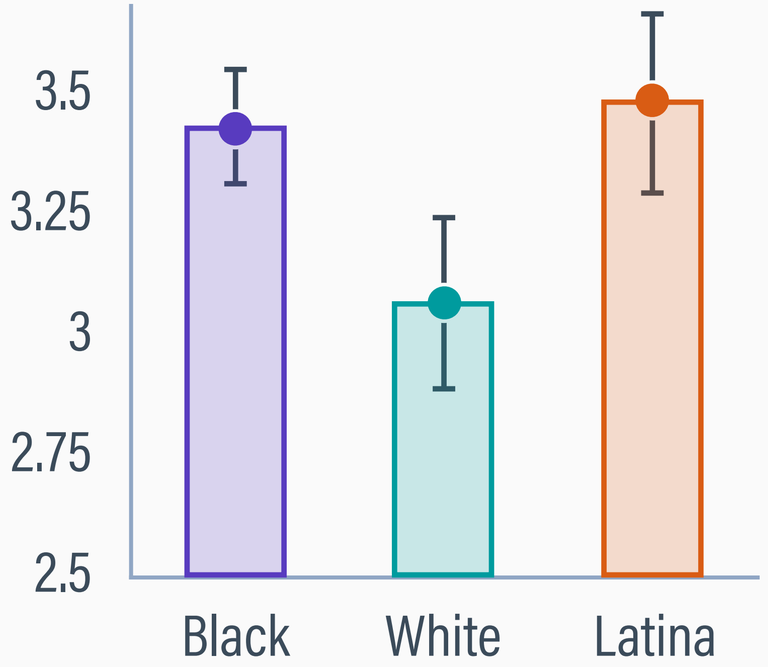
R OUTPUT


Plot Zoom
3.5
3.3
3.1
2.9
AnxiousArousal
Black
Latina
White
Race

ANALYSIS SUMMARY

Group	$\bar{X}$	s	n
Black	3.43	1.96	901
White	3.07	1.56	389
Latina	3.49	1.61	373

- The mean difference between Black and Latina women is virtually nil (3.49 vs. 3.43, d = 0.03)
- There are small differences between White and Black women (3.07 vs. 3.43, d = 0.20) and White and Latina women (3.07 vs. 3.49, d = 0.23)



SIGNIFICANCE TESTING STEPS

1. Specify hypotheses
2. Define standard of evidence
3. Design study and collect data
4. **Compare data to null hypothesis**
5. Evaluate hypotheses and draw conclusion

COMPARING DATA TO THE NULL

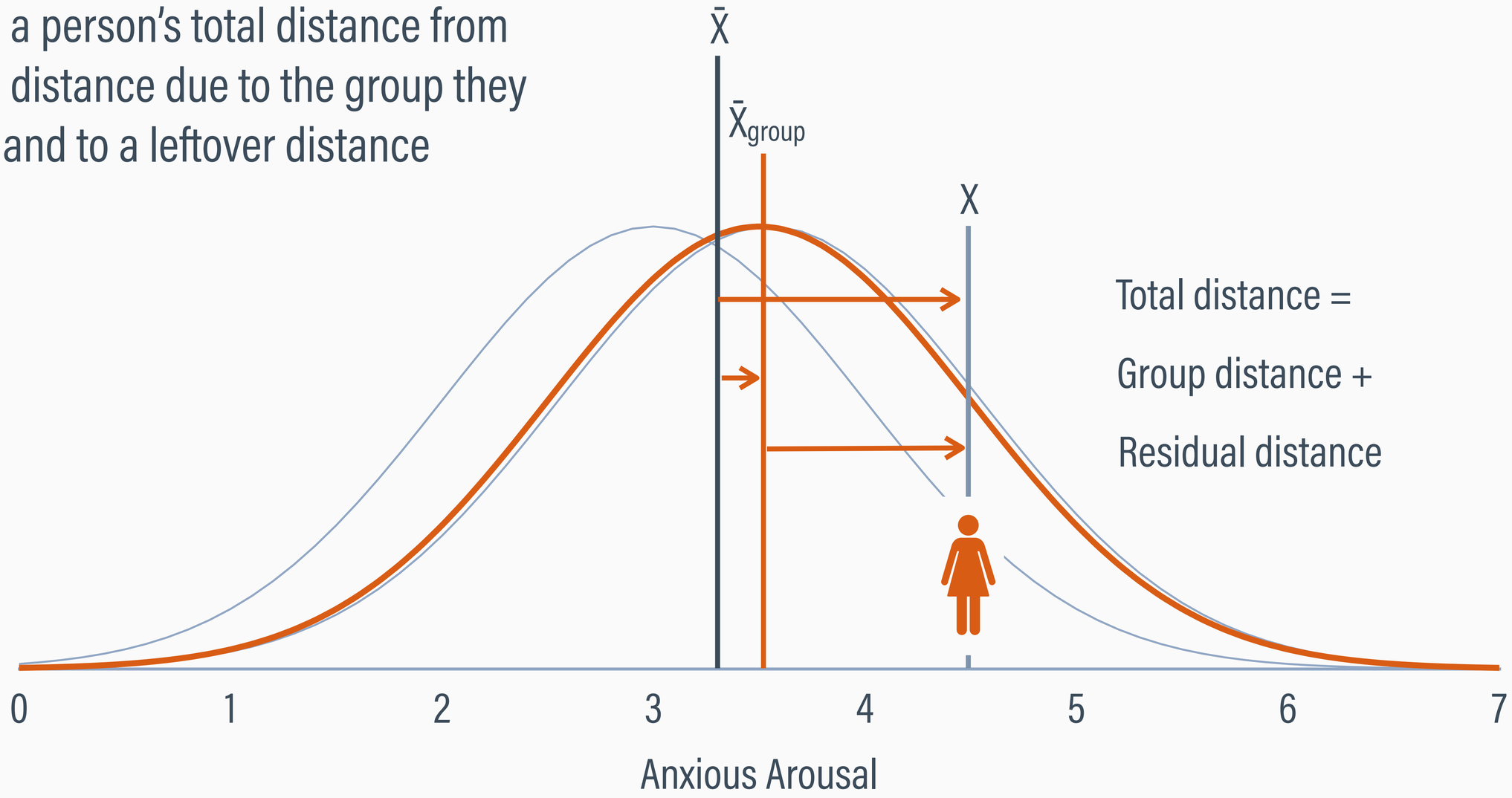
- Two ways to determine whether the sample means are consistent (or inconsistent) with the null population means
- The F-statistic gives a standardized distance between the sample mean differences and the null hypothesis
- A p-value tells us how likely it is that hypothetical samples like our data would originate from the null population

PARTIONING (SEPARATING) VARIATION

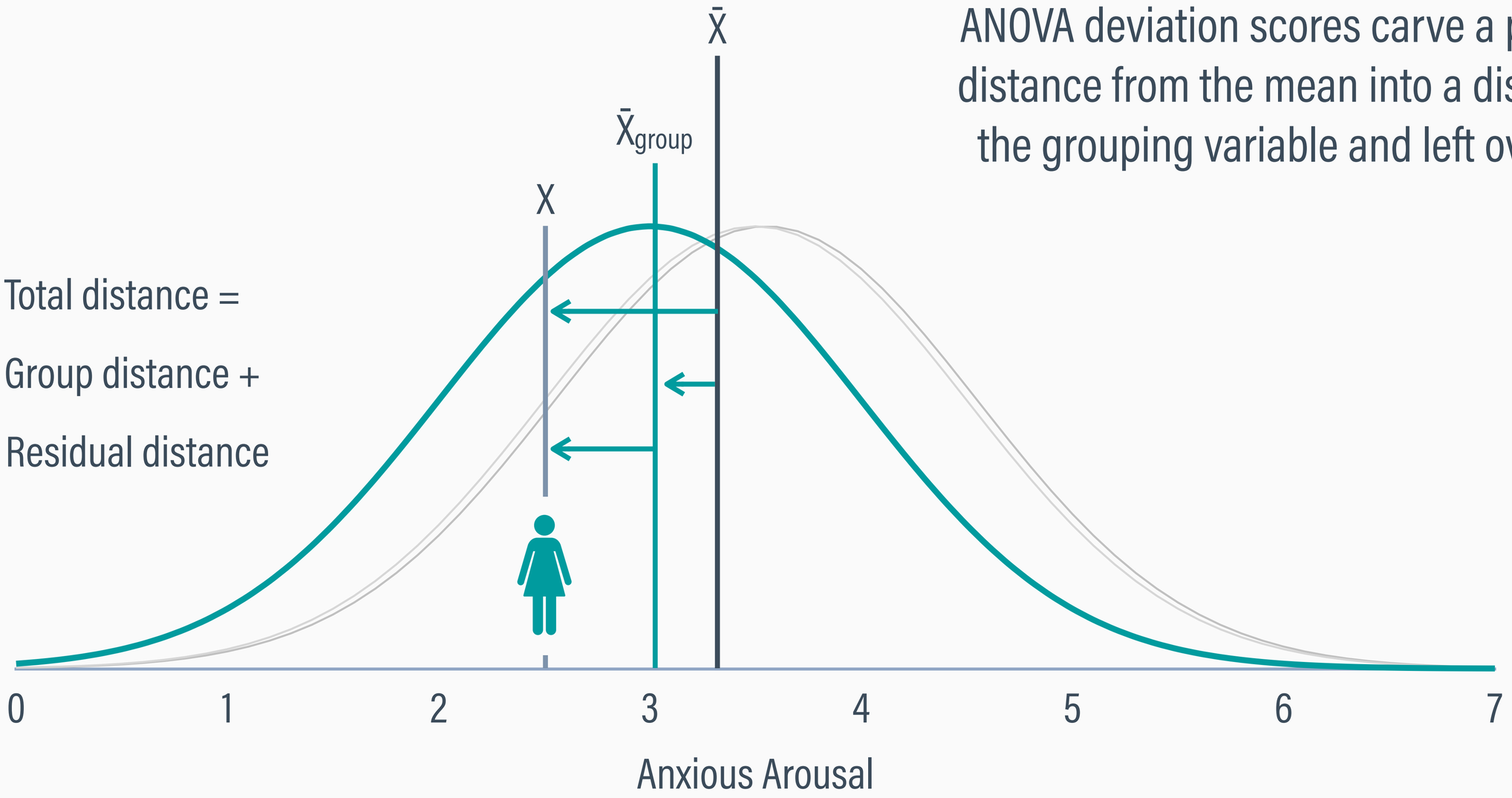
- ANOVA calculates the total variation (e.g., PTSD differences) in the data, then it separates (partitions) variation into two sources
- Differences among the group means capture variation due to the independent variable, and score differences that remain after accounting for group differences reflect leftover variation
- Sum of squares and variances (components of the standard deviation formula) quantity the amount of variation in each

PARTITIONING DISTANCES

ANOVA divides a person's total distance from the mean into a distance due to the group they belong to and to a leftover distance



RESIDUAL (LEFTOVER) DISTANCE



ANOVA deviation scores carve a person's total distance from the mean into a distance due to the grouping variable and left over distance

DEMONSTRATION DATA

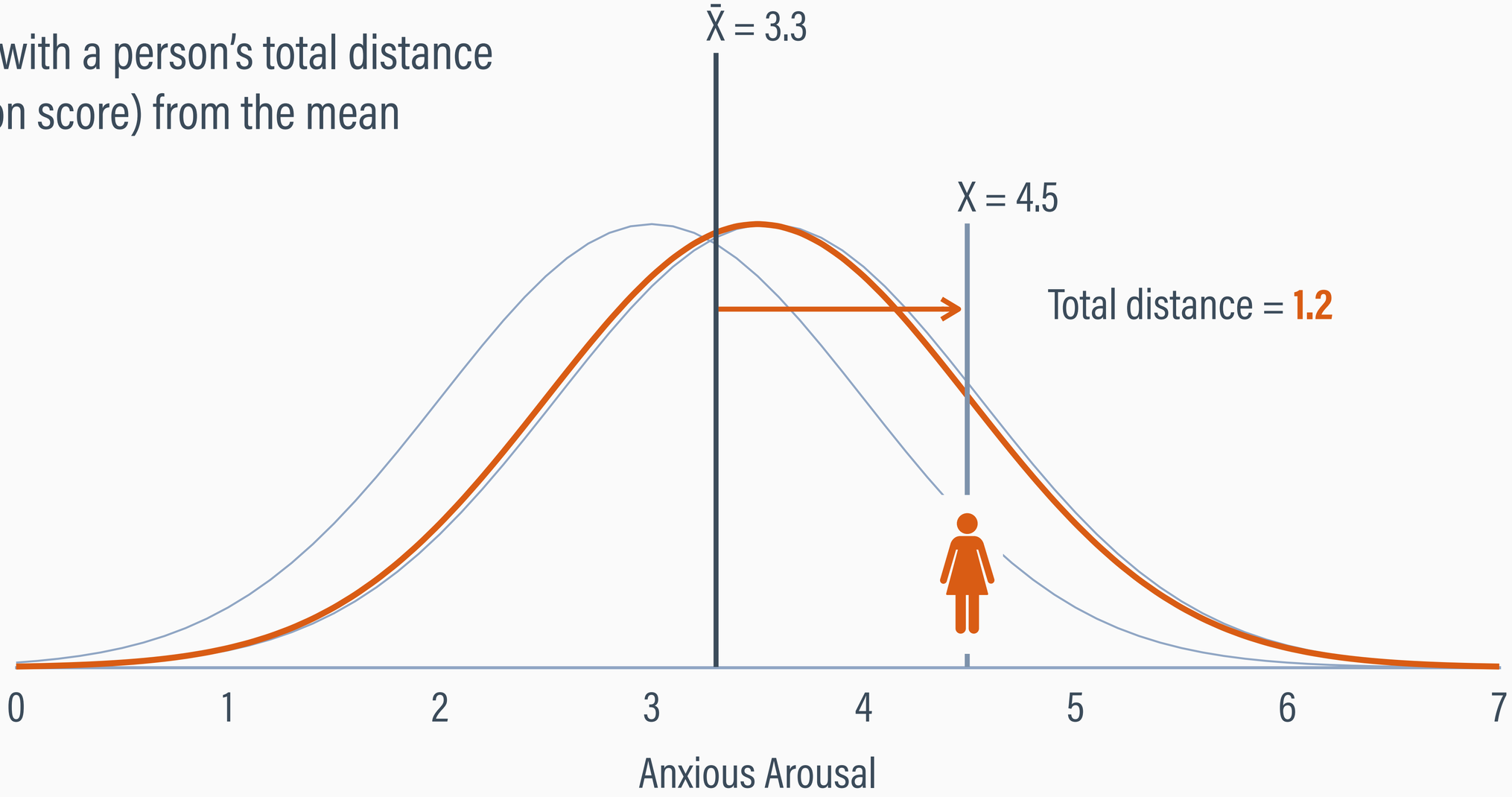
- I demonstrate ANOVA computations using a small subset of data with similar means and standard deviations as the full data set

Group	$\bar{X}_{group}$	s	n
Black	3.40	2.30	5
White	3.00	1.73	5
Latina	3.60	1.82	5

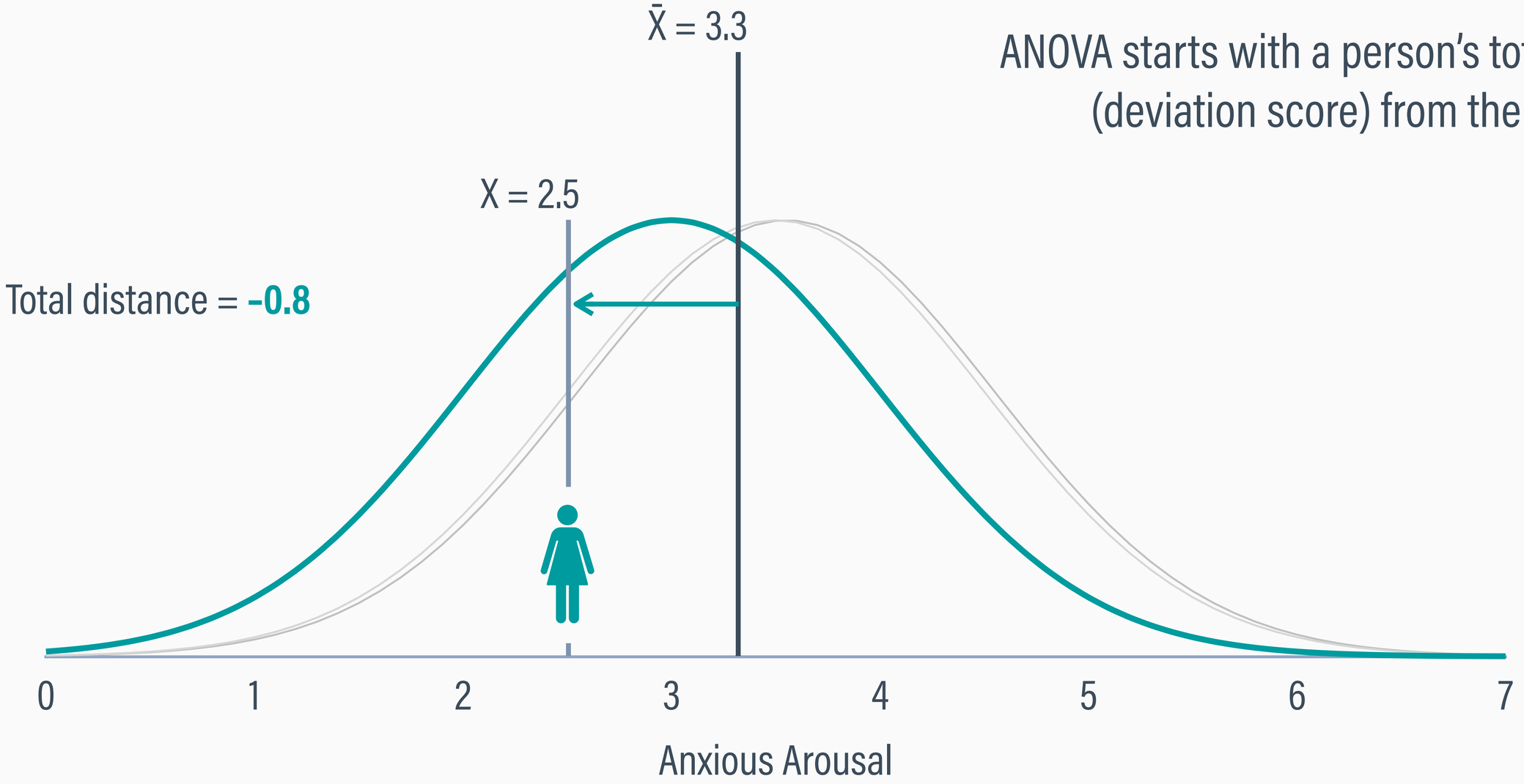
Group	X	$\bar{X}_{group}$	$\bar{X}$
Black	1	3.4	3.3
Black	2	3.4	3.3
Black	3	3.4	3.3
Black	4	3.4	3.3
Black	5	3.4	3.3
White	6	3.0	3.3
White	2	3.0	3.3
White	2	3.0	3.3
White	2	3.0	3.3
White	3	3.0	3.3
Latina	1	3.6	3.3
Latina	3	3.6	3.3
Latina	4	3.6	3.3
Latina	4	3.6	3.3
Latina	6	3.6	3.3

TOTAL DISTANCE (DEVIATION SCORES)

ANOVA starts with a person's total distance (deviation score) from the mean



TOTAL DISTANCE (DEVIATION SCORES)



ANOVA starts with a person's total distance (deviation score) from the mean

SUM OF SQUARES REVIEWED

- Summing the squared distances from the scores to the overall mean (called the grand mean in ANOVA) gives a foundational measure called sum of squares (SS)

$$SS_{total} = \sum(X - \bar{X})^2 = \sum(\text{score} - \text{grand mean})^2$$

- The sum of squares expresses the total amount of variability in the data as a lump sum

TOTAL SUM OF SQUARES

$$SS_{total} = \sum(X - \bar{X})^2 + \sum(X - \bar{X})^2 + \sum(X - \bar{X})^2 = 21.22 + 12.54 + 13.56 = 47.33$$

Black
(Group 1)

X	$\bar{X}_{group}$	$\bar{X}$	d	d^2
1	3.4	3.3	-2.33	5.43
2	3.4	3.3	-1.33	1.77
3	3.4	3.3	-0.33	0.11
4	3.4	3.3	0.67	0.45
7	3.4	3.3	3.67	13.47

Sum of squares = **21.22**

White
(Group 2)

X	$\bar{X}_{group}$	$\bar{X}$	d	d^2
6	3.0	3.3	2.67	7.13
2	3.0	3.3	-1.33	1.77
2	3.0	3.3	-1.33	1.77
2	3.0	3.3	-1.33	1.77
3	3.0	3.3	-0.33	0.11

Sum of squares = **12.54**

Latina
(Group 3)

X	$\bar{X}_{group}$	$\bar{X}$	d	d^2
1	3.6	3.3	-2.33	5.43
3	3.6	3.3	-0.33	0.11
4	3.6	3.3	0.67	0.45
4	3.6	3.3	0.67	0.45
6	3.6	3.3	2.67	7.13

Sum of squares = **13.56**

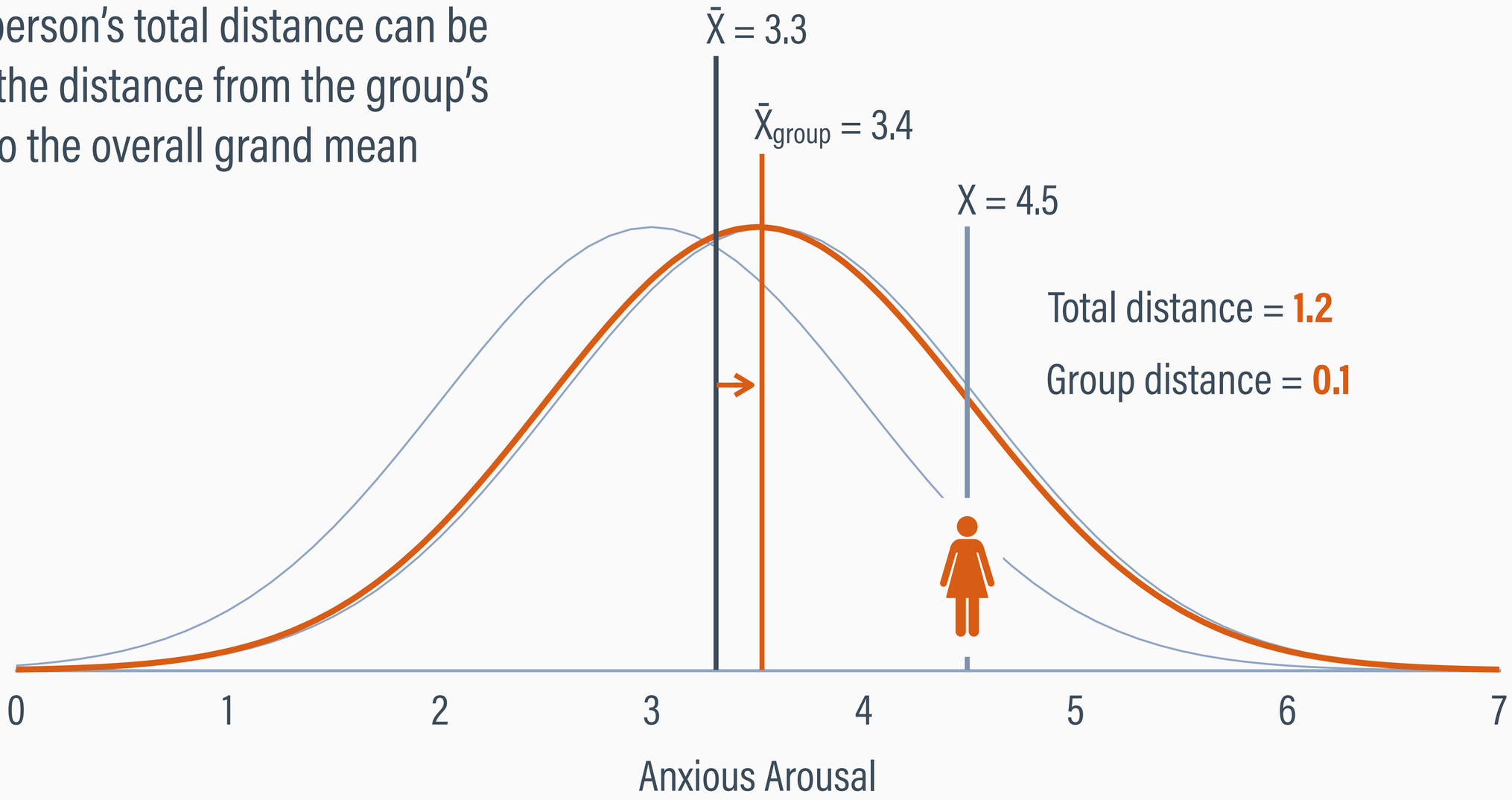
TOTAL PTSD VARIABILITY

- Represented as a lump sum, the total variation among PTSD scores is SS = 47.33
- Next, modified sum of squares formulas will divide the variability "pie" into slices attributable to the independent variable and to leftover sources

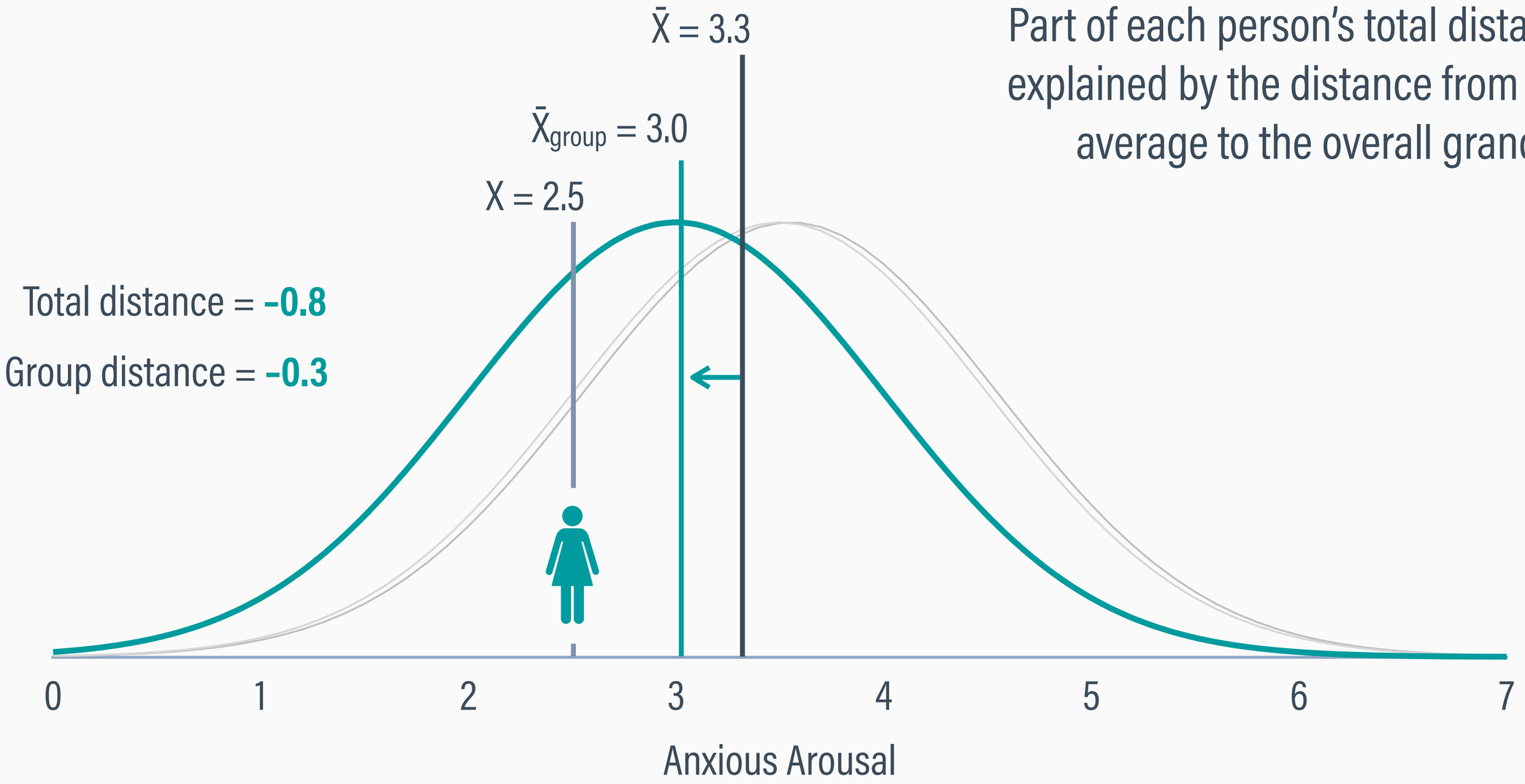
Sum of squares groups =	?
Sum of squares residual =	?
Sum of squares total =	**47.33**

GROUP DISTANCE (MEAN DIFFERENCE)

Part of each person's total distance can be explained by the distance from the group's average to the overall grand mean



GROUP DISTANCE (MEAN DIFFERENCE)



Part of each person's total distance can be explained by the distance from the group's average to the overall grand mean

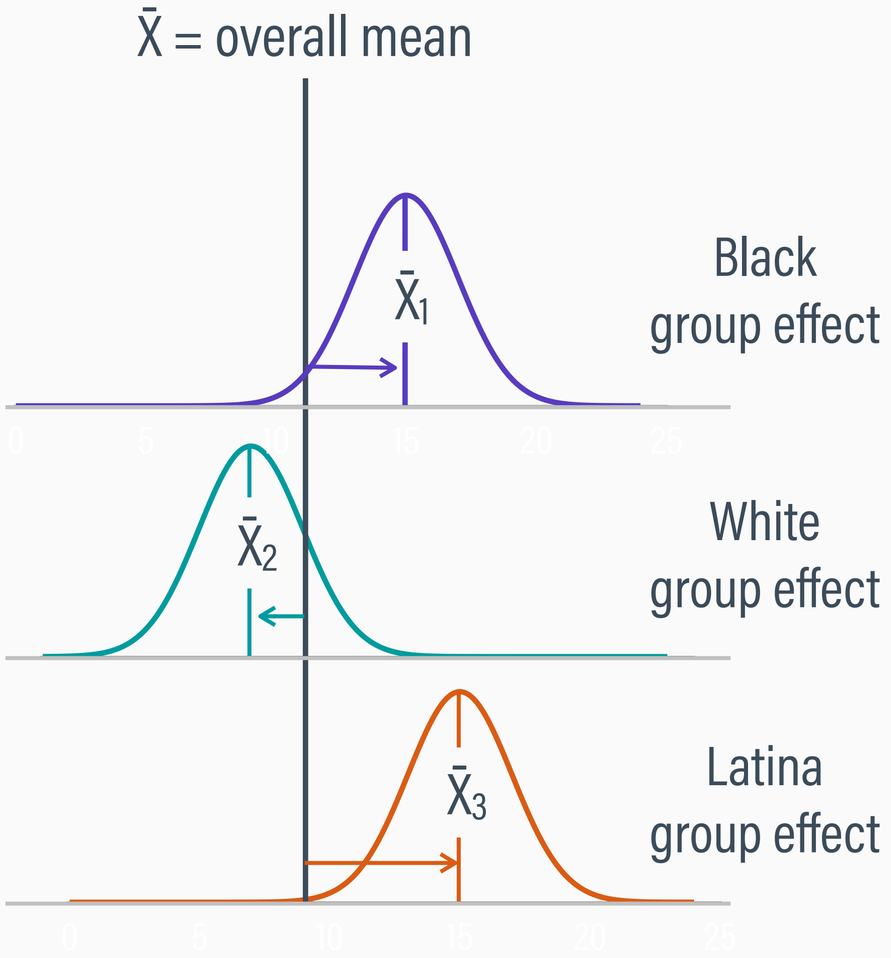
GROUP SUM OF SQUARES

- Variation due the grouping (independent) variable involves distances from the group-specific averages to the overall mean

$$SS_{group} = \sum(\bar{X}_{group} - \bar{X})^2$$

$$= \sum(\text{group mean} - \text{grand mean})^2$$

- The sum of squares formula uses a distance (deviation) score comparing these two means



GROUP SUM OF SQUARES

$$SS_{group} = \sum(\bar{X}_{group} - \bar{X})^2 + \sum(\bar{X}_{group} - \bar{X})^2 + \sum(\bar{X}_{group} - \bar{X})^2 = 0.02 + 0.54 + 0.46 = 0.93$$

Black
(Group 1)

X	$\bar{X}_{group}$	$\bar{X}$	d	d^2
1	3.4	3.33	0.07	0.00
2	3.4	3.33	0.07	0.00
3	3.4	3.33	0.07	0.00
4	3.4	3.33	0.07	0.00
7	3.4	3.33	0.07	0.00

Sum of squares = **0.02**

White
(Group 2)

X	$\bar{X}_{group}$	$\bar{X}$	d	d^2
6	3.0	3.33	-0.33	0.11
2	3.0	3.33	-0.33	0.11
2	3.0	3.33	-0.33	0.11
2	3.0	3.33	-0.33	0.11
3	3.0	3.33	-0.33	0.11

Sum of squares = **0.54**

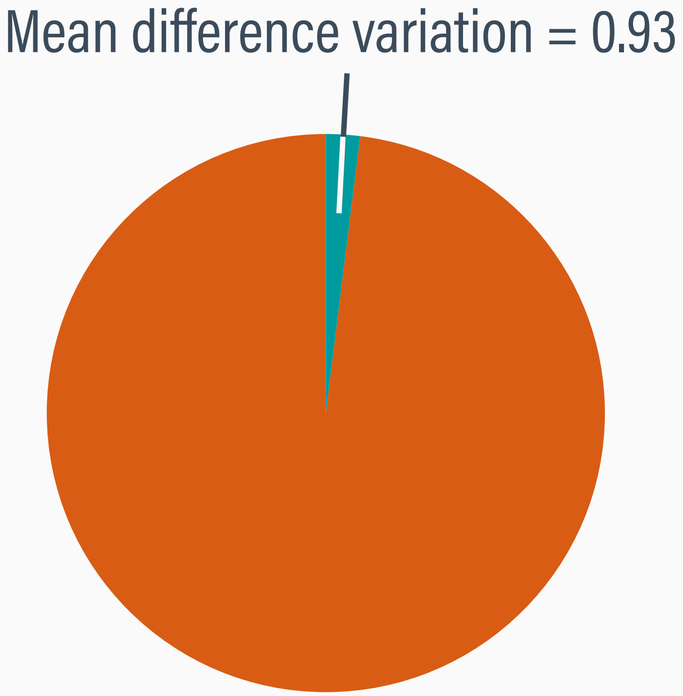
Latina
(Group 3)

X	$\bar{X}_{group}$	$\bar{X}$	d	d^2
1	3.6	3.33	0.27	0.07
3	3.6	3.33	0.27	0.07
4	3.6	3.33	0.27	0.07
4	3.6	3.33	0.27	0.07
6	3.6	3.33	0.27	0.07

Sum of squares = **0.36**

MEAN DIFFERENCE VARIABILITY

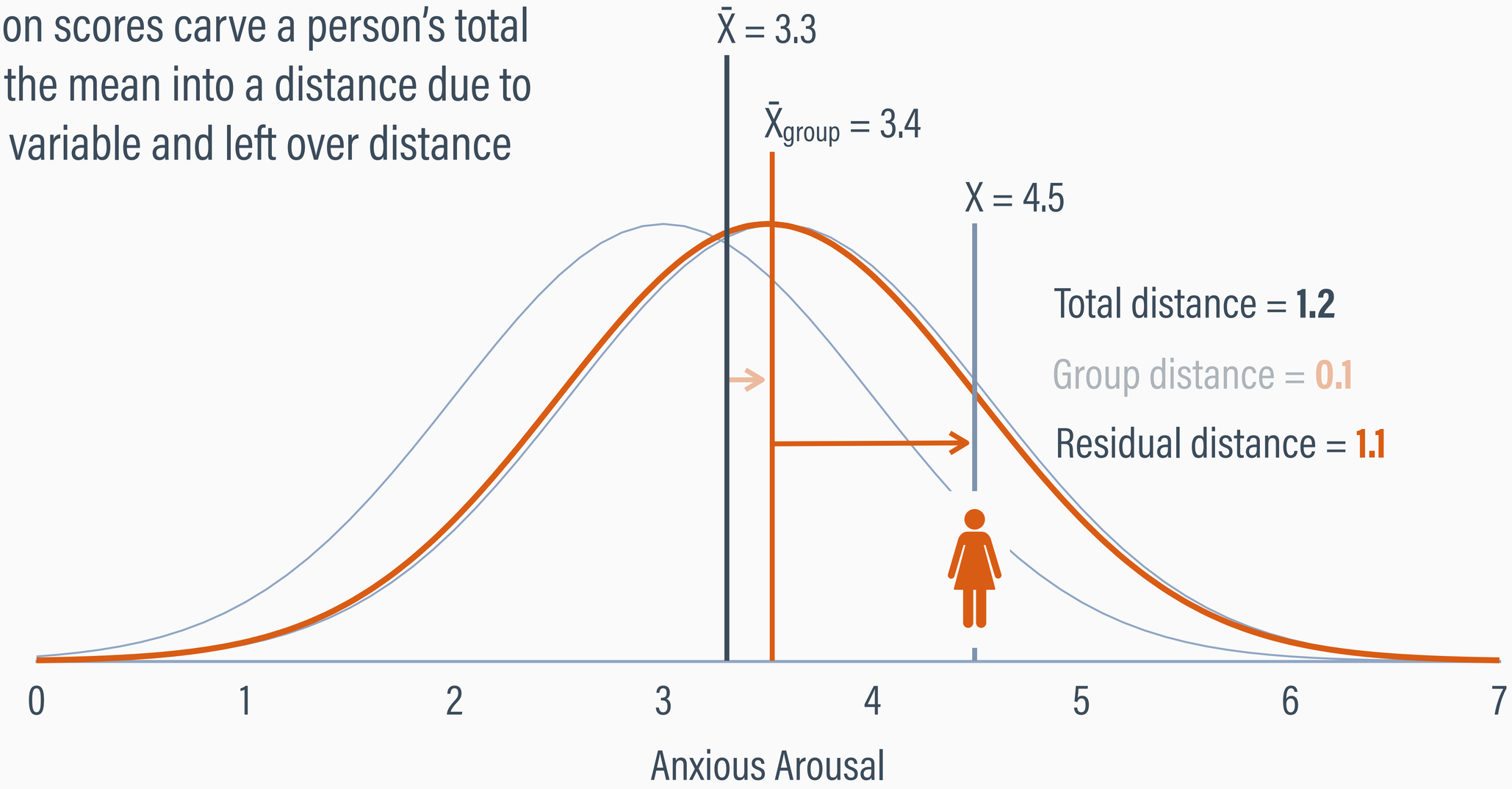
- Represented as a lump sum, the PTSD variation due to racial groups is $SS_{group} = 0.93$
- A portion of the total PTSD variation ($SS_{total} = 47.33$) is captured by which group people belong to (the independent variable)



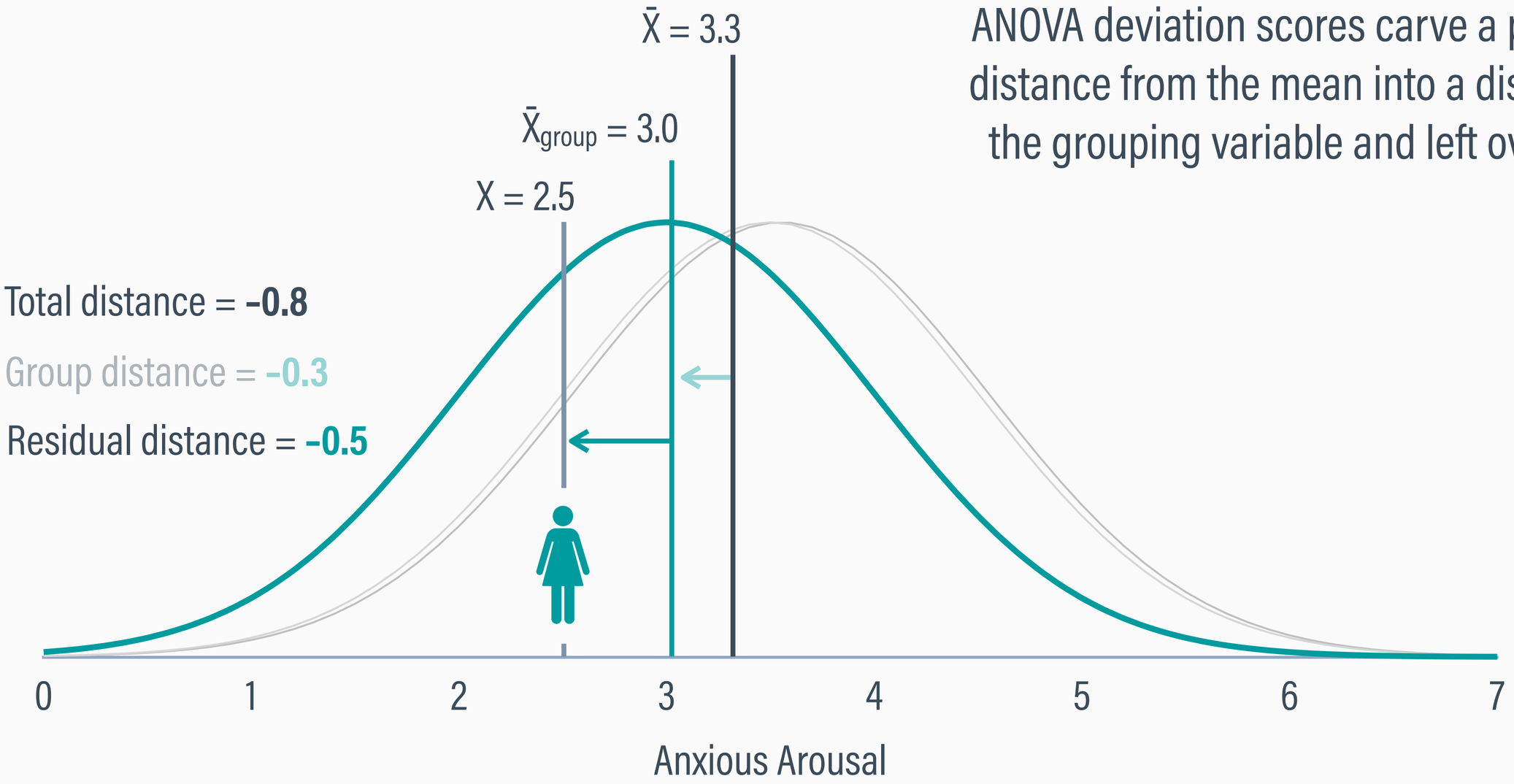
Sum of squares groups =	**0.93**
Sum of squares residual =	**?**
Sum of squares total =	**47.33**

RESIDUAL (LEFTOVER) DISTANCE

ANOVA deviation scores carve a person's total distance from the mean into a distance due to the grouping variable and left over distance



RESIDUAL (LEFTOVER) DISTANCE



ANOVA deviation scores carve a person's total distance from the mean into a distance due to the grouping variable and left over distance

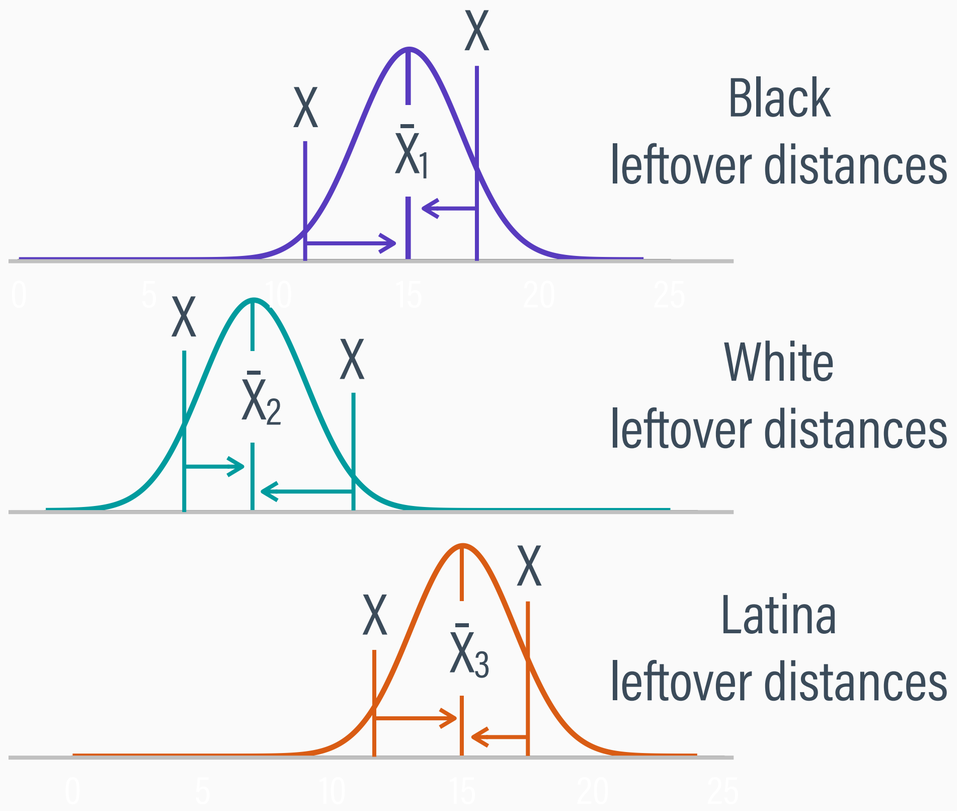
RESIDUAL SUM OF SQUARES

- Residual variation reflects the leftover part of the scores that is not attributable to group differences

$$SS_{residual} = \sum (X - \bar{X}_{group})^2$$

$$= \sum (\text{score} - \text{group mean})^2$$

- The sum of squares formula uses a distance that compares scores to their group-specific averages



RESIDUAL SUM OF SQUARES

$$SS_{residual} = \sum(X - \bar{X}_{group})^2 + \sum(X - \bar{X}_{group})^2 + \sum(X - \bar{X}_{group})^2 = 21.20 + 12.00 + 13.20 = 46.40$$

Black
(Group 1)

X	$\bar{X}_{group}$	$\bar{X}$	d	d^2
1	3.4	3.3	-2.40	5.76
2	3.4	3.3	-1.40	1.96
3	3.4	3.3	-0.40	0.16
4	3.4	3.3	0.60	0.36
7	3.4	3.3	3.60	12.96

Sum of squares = **21.20**

White
(Group 2)

X	$\bar{X}_{group}$	$\bar{X}$	d	d^2
6	3.0	3.3	3.00	9.00
2	3.0	3.3	-1.00	1.00
2	3.0	3.3	-1.00	1.00
2	3.0	3.3	-1.00	1.00
3	3.0	3.3	0.00	0.00

Sum of squares = **12.00**

Latina
(Group 3)

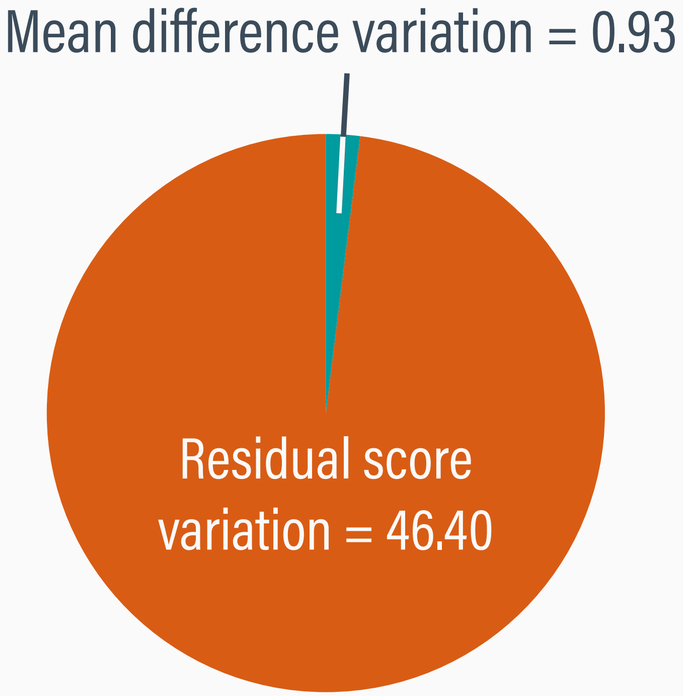
X	$\bar{X}_{group}$	$\bar{X}$	d	d^2
1	3.6	3.3	-2.60	6.76
3	3.6	3.3	-0.60	0.36
4	3.6	3.3	0.40	0.16
4	3.6	3.3	0.40	0.16
6	3.6	3.3	2.40	5.76

Sum of squares = **13.20**

PARTITIONED VARIATION

- The ANOVA sum of squares formulas divide the total variation in the data into two non-overlapping sources

- Score variation due to the independent variable (group mean differences), and left over (residual) variation



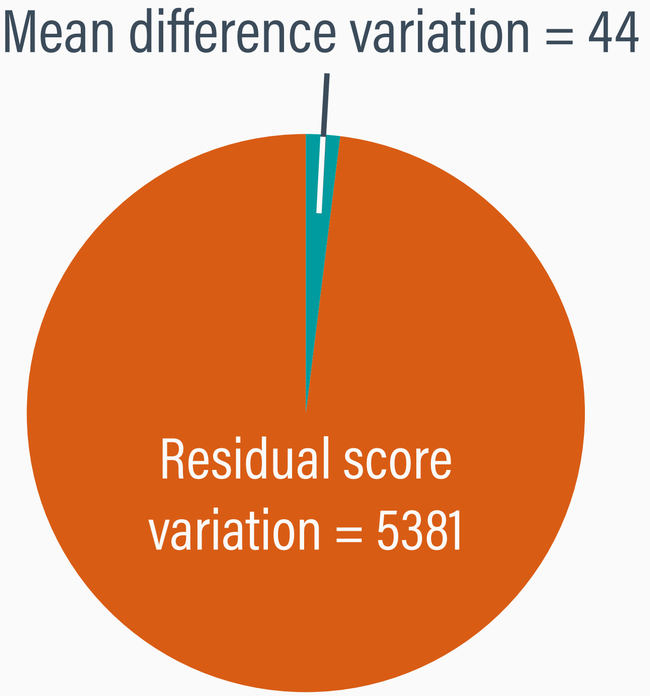
Sum of squares groups =	**0.93**
Sum of squares residual =	**46.40**
Sum of squares total =	**47.33**

NEXT STEPS: EFFECT SIZE AND F-STATISTIC

- The sum of squares values are building blocks of ANOVA's effect size measure (practical significance) and F-statistic (statistical significance)
- The R^2 effect size (also called η^2) is the proportion of the total variability "pie" that is attributable to group mean differences (the independent variable)
- The F-statistic is a ratio comparing group mean variation to residual variation (i.e., a signal-to-noise ratio)

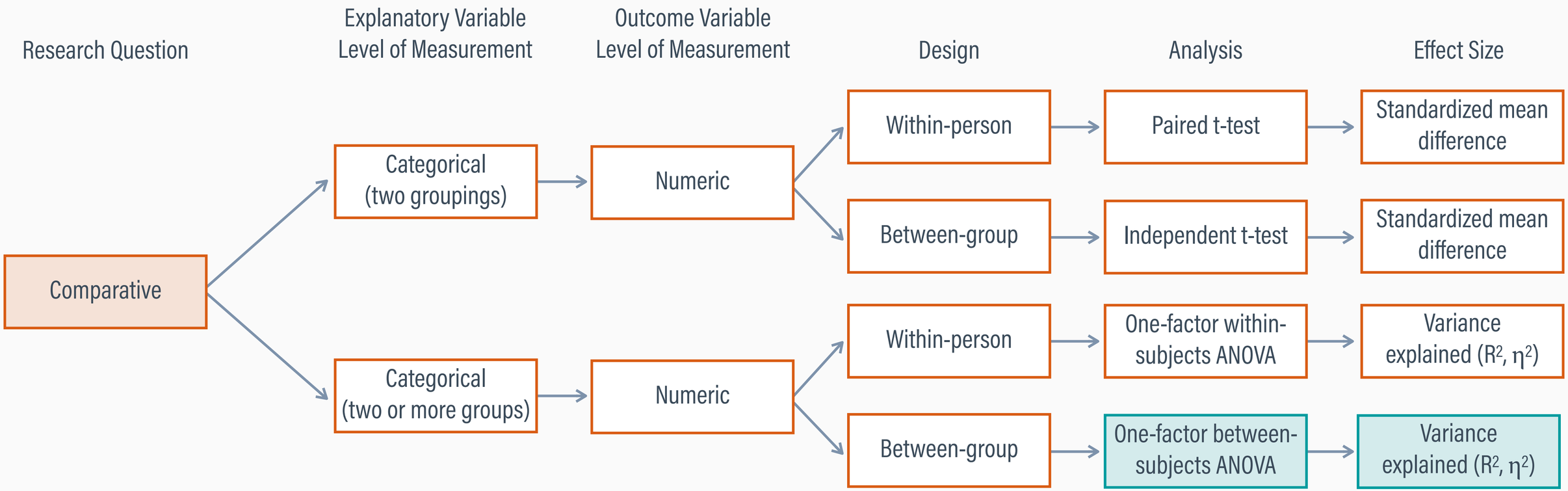
PARTITIONED VARIATION (FULL SAMPLE)

- The ANOVA sum of squares formulas divide the total variation in the data into two non-overlapping sources
- Score variation due to the independent variable (group mean differences), and left over (residual) variation



Sum of squares groups =	**44**
Sum of squares residual =	**5381**
Sum of squares total =	**5425**

STATISTICAL ORG CHART

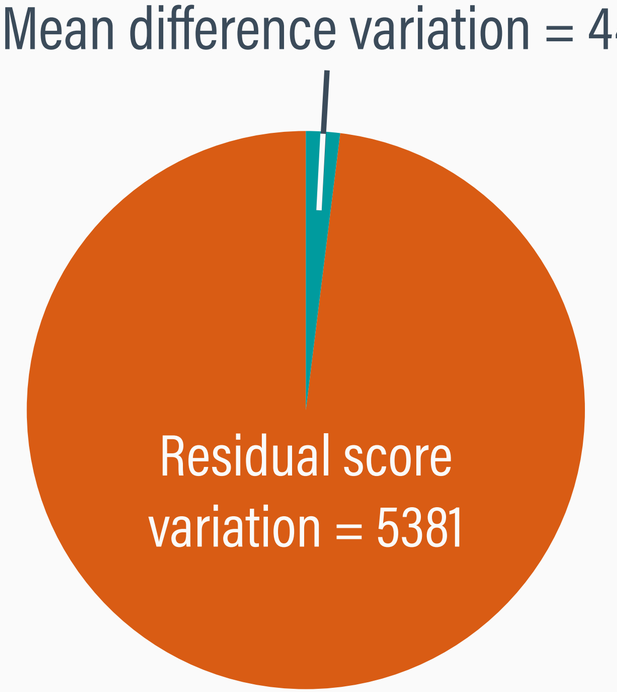


R-SQUARE EFFECT SIZE

- Racial/ethnic mean differences account for about 1% of the total PTSD variation

$$R^2 = \eta^2 = \frac{SS_{group}}{SS_{total}} = \frac{44}{5425} = .01$$

- Score differences attributable to the grouping (independent) variable comprise about 1% of the total variation pie



Sum of squares groups =	**44**
Sum of squares residual =	**5381**
Sum of squares total =	**5425**

R OUTPUT

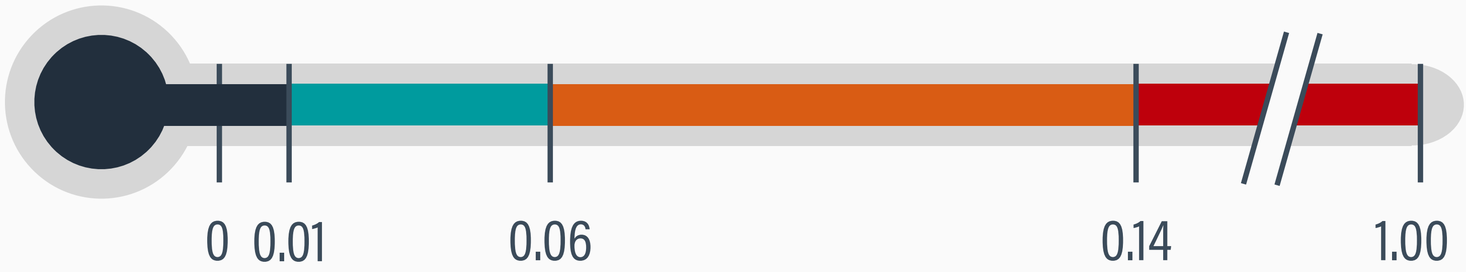
```
# R2 for Anova Regression
       R2: 0.008
  adj. R2: 0.007
```

R-SQUARE EFFECT SIZE GUIDELINES

Small = .01 to .06 (1% to 6%)

Moderate = .06 to .14 (6% to 14%)

Large = greater than .14 (14% to 100%)



VARIANCE REVIEWED

- The variance is the averaged squared distance between the scores and the mean

$$s^2 = \frac{\sum(X - \bar{X})^2}{N - 1} = \frac{\text{Sum of squares}}{\text{Adjusted sample size (df)}}$$

- The degrees of freedom (N – 1) is the adjusted sample size needed to get an unbiased estimate of variation

MEAN SQUARES

- In ANOVA, the averaged squared distance is called a mean square instead of a variance
- The mean square for the grouping (independent) variable is an average squared distance per group
- The residual mean square is an averaged squared distance per person

$$MS_{group} = \frac{SS_{group}}{G - 1}$$

$$= \frac{\text{Sum of squared distances}}{\text{Adjusted number of groups } (df_{group})} = \frac{0.93}{3 - 1}$$

$$= 0.47$$

$$MS_{residual} = \frac{SS_{residual}}{N - G}$$

$$= \frac{\text{Sum of squared distances}}{\text{Adjusted sample size } (df_{residual})} = \frac{46.60}{15 - 3}$$

$$= 3.87$$

ANOVA SUMMARY TABLE

- The two sources of variation are displayed in an ANOVA summary table
- The sum of squares (SS) values reflect variation as a lump sum, and the mean squares (MS) values give variation as an average squared distance

Mean squares = variability as an averaged squared distance

Sum of squares = variability as a sum of squared distances

Source	SS	df	MS
Group (Race)	44.06	2	22.03
Residual	5380.62	1660	3.24

Degrees of freedom = adjusted number of groups and participants

R OUTPUT

```
               Df Sum Sq Mean Sq F value  Pr(>F)
Race            2     44  22.032   6.797 0.00115 **
Residuals    1660   5381   3.241
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

F-STATISTIC

- The F-statistic is a ratio comparing group mean variation to residual variation (i.e., a signal-to-noise ratio)

$$F = \frac{MS_{group}}{MS_{residual}} = \frac{\text{Group mean difference variation}}{\text{Left over individual variation}}$$

- If the null is true, the F-statistic should be close to 1

F-STATISTIC EXAMPLE

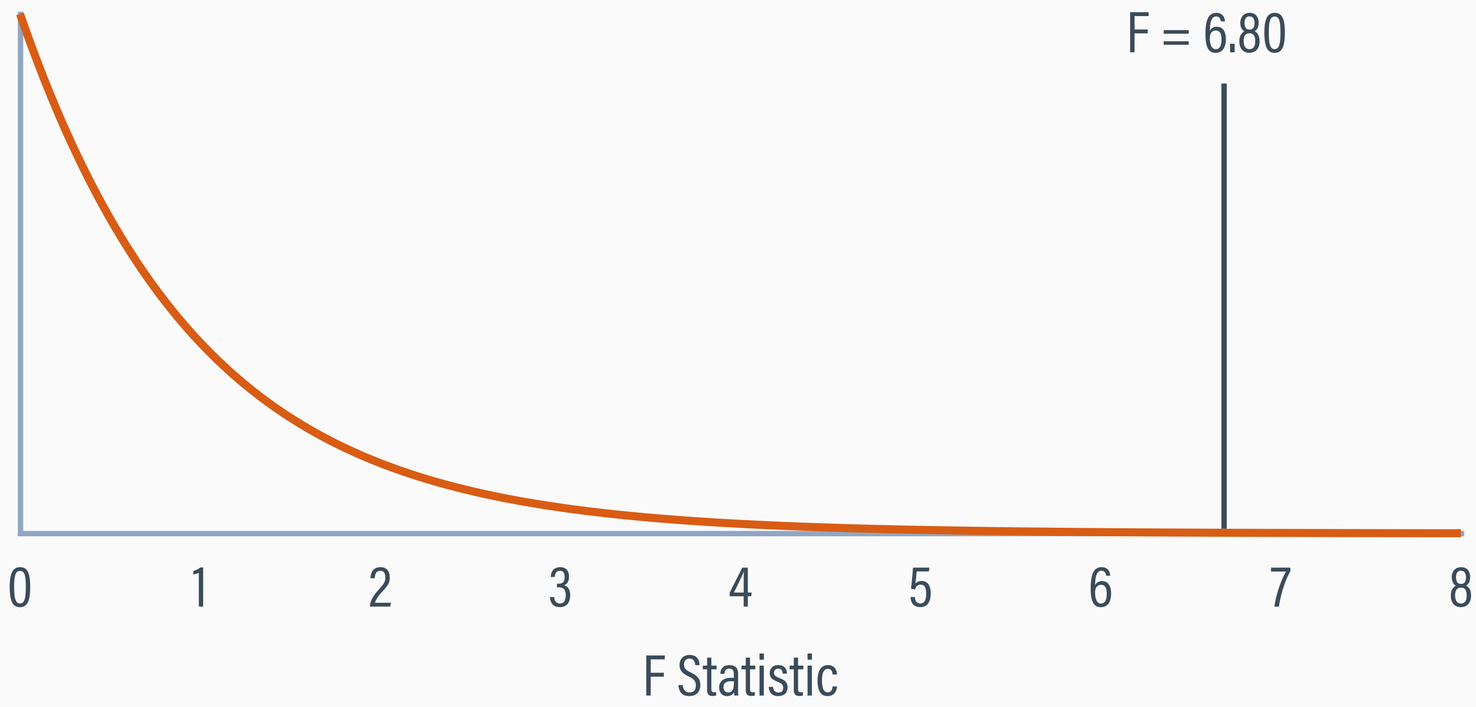
- The F-statistic compares variation explained by the independent variable to unexplained (error) variation

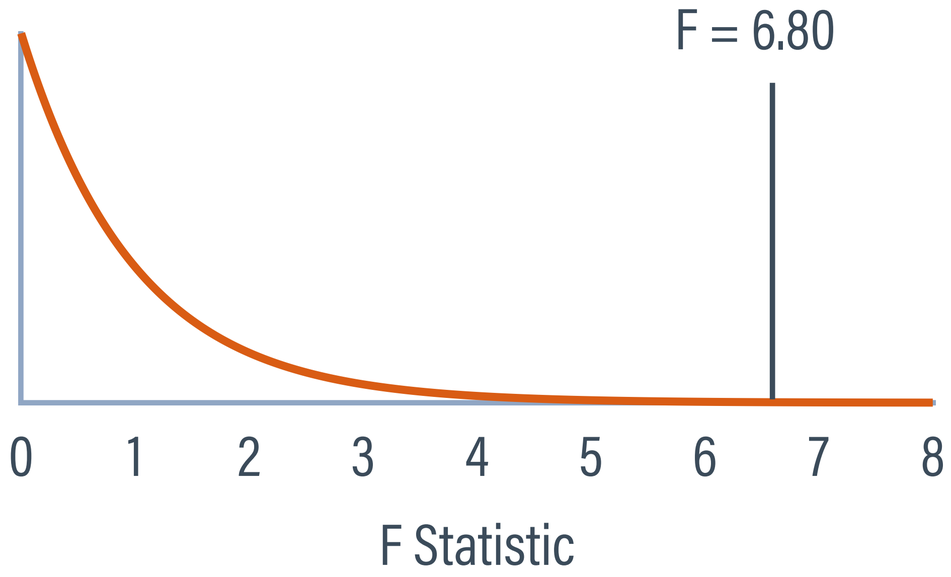
$$F = \frac{MS_{group}}{MS_{residual}} = \frac{22.03}{3.24} = 6.80$$

- Variation due to racial group mean differences is about 6.8 times larger than leftover PTSD variation within groups

SAMPLING DISTRIBUTION FOR PTSD ANALYSIS

- F distribution for an analysis with three groups ($df_{group} = 2$) and 1663 participants ($df_{residual} = 1660$)





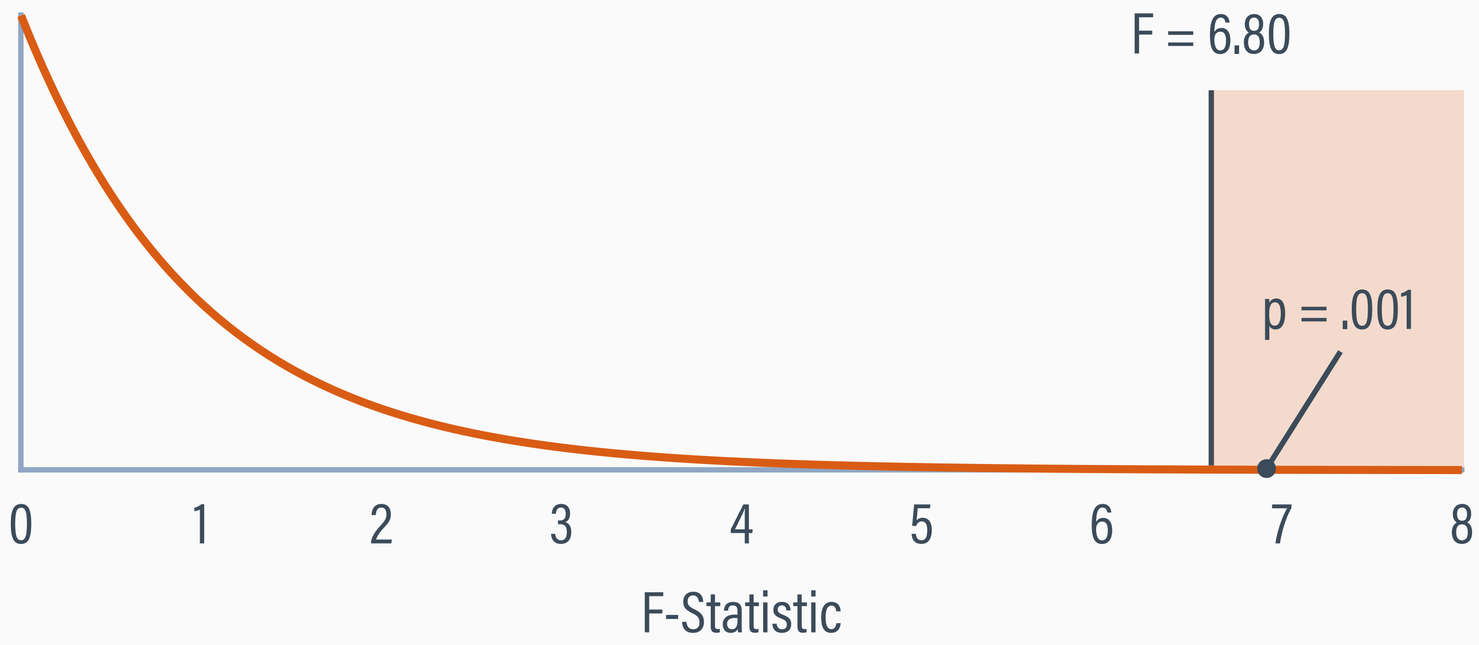
The F-statistic for the example was 6.81. In small groups of two or three, explain what the magnitude of the F-statistic indicates about the plausibility of the null hypothesis. By eyeballing the sampling distribution, estimate the probability of obtaining an F-statistic as large as 6.8 from a null population where all groups are equal.

PROBABILITY VALUES (P-VALUES)

- A p-value is defined as proportion of hypothetical samples that have a F-statistic at least as large as the sample data
- Assuming the null is true, how likely is it to draw a sample with an effect at least as large as the one from our data?
- Visually, probability is an area under the curve, obtained by applying calculus integrals to the F-distribution function

EXACT PROBABILITY

- If the null is true in the population, 1 out of every 1000 hypothetical samples would produce F-statistics ≥ 6.80 (mean differences as large as those observed in this sample)



R OUTPUT

```
               Df Sum Sq Mean Sq F value  Pr(>F)
Race            2     44  22.032   6.797 0.00115 **
Residuals    1660   5381   3.241
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

SIGNIFICANCE TESTING STEPS

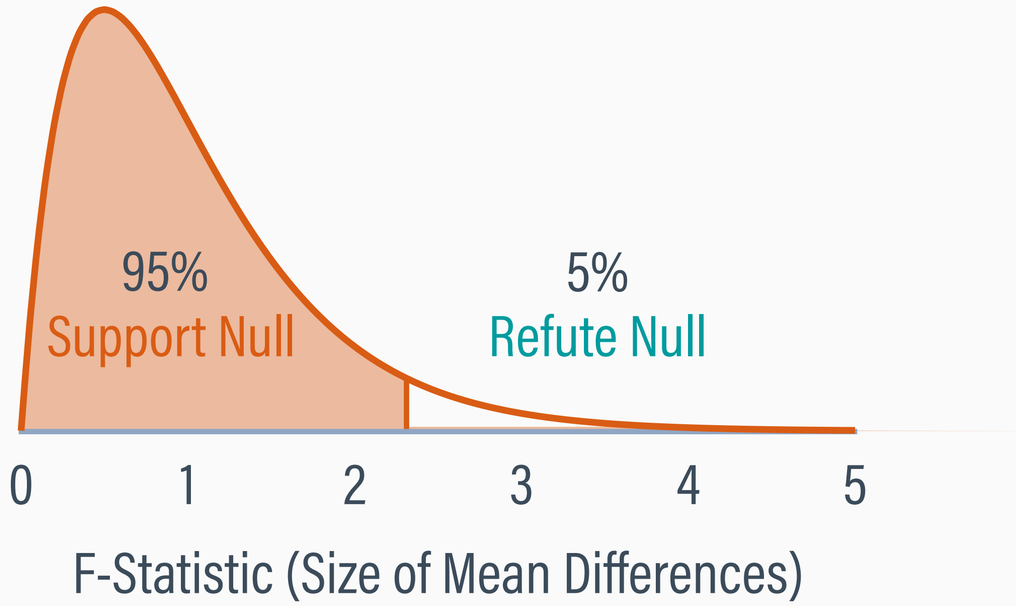
1. Specify hypotheses
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

RESEARCH QUESTION REVISITED

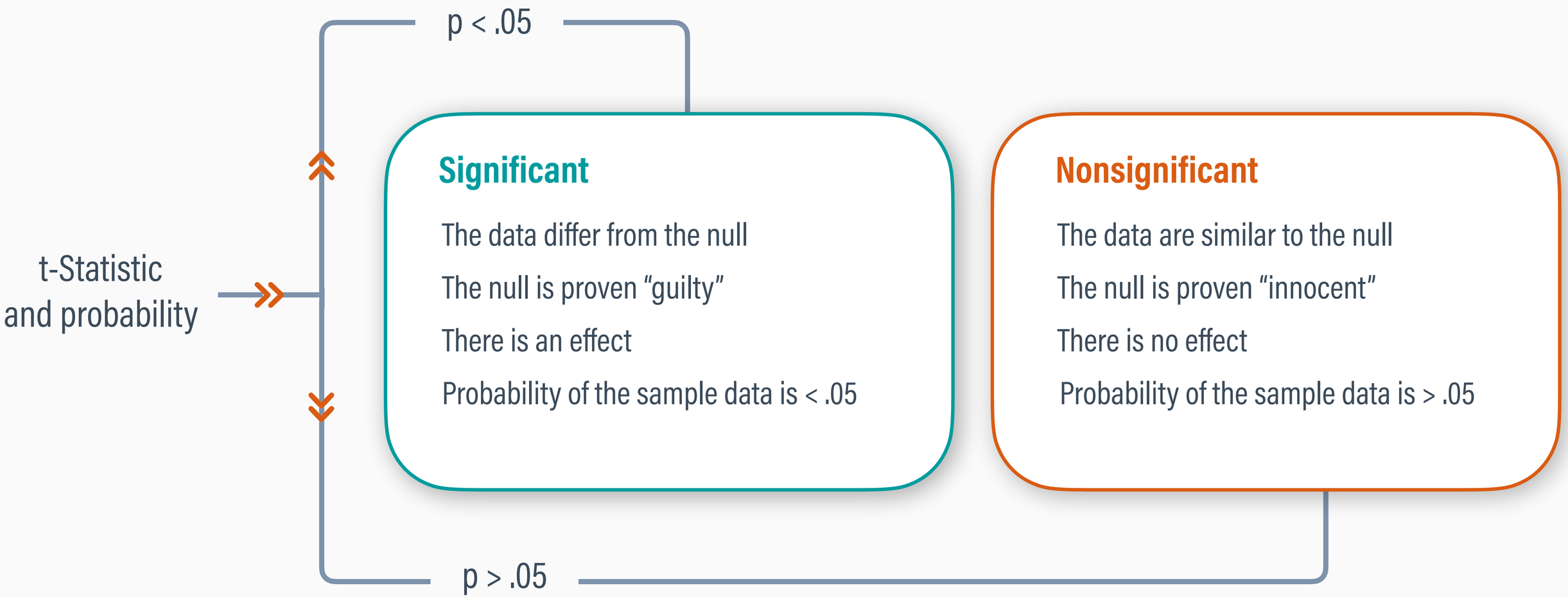
- Studies typically attempt to answer a handful of research questions involving associations between key variables
- Do women of color (Black and Latina) differ from White women in their levels of PTSD anxious arousal symptoms?
- The null (no effect) hypothesis states that the three groups are identical (all population mean differences are zero)

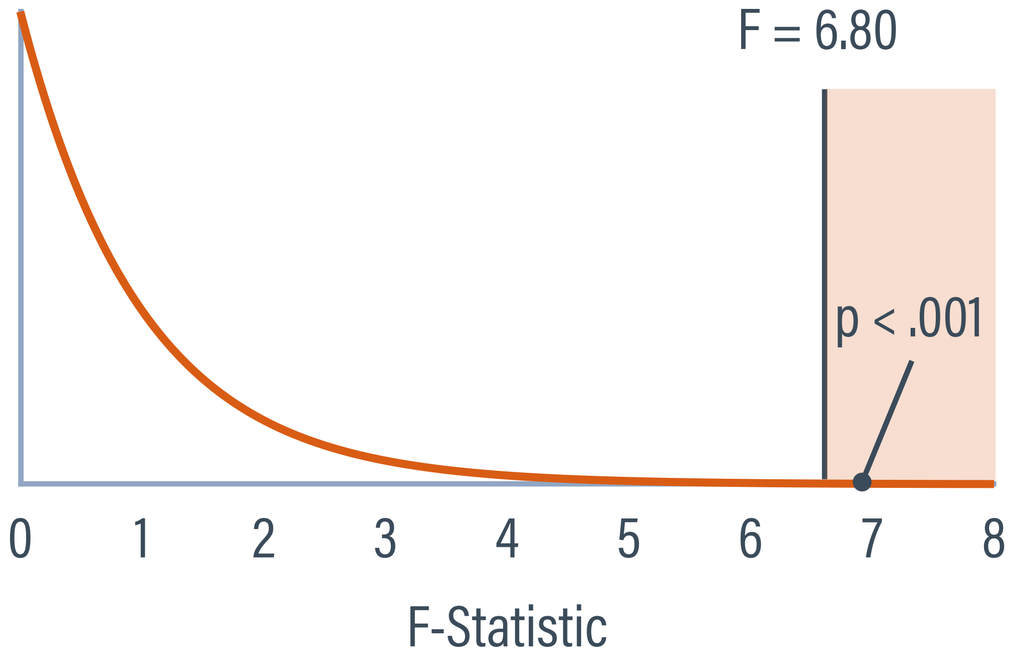
5% REJECTION REGION REVISITED

- By convention, we refute the null if the sample mean differences (encoded by an F-statistic) fall in the most extreme 5% of the sampling distribution
- Such as sample has less than a 5% chance of originating from the null population ($p < .05$)
- We deem the null implausible because our data are unlikely to originate from that population



DECISION TREE





The probability for the F-statistic is $p < .001$ (less than 1 out of 1000). In small groups of two or three, discuss your decision about the the null hypothesis. Translate your decision into a tangible statement about race/ethnic group differences in PTSD experiences.

CONCLUSION SO FAR

- A p-value of $< .001$ provides evidence against the null
- Mean differences as large those observed in this study (encoded by $F \geq 6.80$) are unlikely to have originated from a null population where all groups have identical PTSD levels
- The sample data provide evidence that women from different racial groups experience different levels of PTSD

FALSE POSITIVES (TYPE I ERRORS)

- The 5% rejection region is an area of the F distribution that contains outlier samples that are unlikely *but not impossible*
- When F falls in the rejection region (evidence against the null), there is still a 5% chance it came from the null population
- We conclude there are differences, while acknowledging that there is a 5% chance of a false positive—incorrectly rejecting the null when it is actually true (a Type I error)

APA-STYLE WRITE-UP

We performed a one-factor ANOVA analysis to assess anxious arousal PTSD differences among women from three racial/ethnic groups: Black, White, and Latina. Table 1 gives the means and standard deviations for each group. The ANOVA analysis revealed statistically significant differences among the group means, $F(2,1660) = 6.80$, $p < .001$. Further, the R^2 effect size indicated that group differences explained approximately 1% of the variation in PTSD scores, which is a small effect by conventional standards (Cohen, 1988).

APA-STYLE TABLE

- All formatting (underlying, italics, and symbols) are prescribed by the APA Style Manual (7th Edition)

Table 1

Descriptive Statistics by Racial/Ethnic Group

Group	*M*	*SD*	*n*
Black	3.43	1.96	901
White	3.07	1.56	389
Latina	3.49	1.61	373

PRELIMINARY CONCLUSION

- The sample mean differences were “significantly significant”
- It is highly unlikely that this data set originated from a population where all three groups are identical
- The sample data provide evidence that women from different racial groups experience different levels of PTSD

OUTLINE

1. ANOVA overview
2. Significance testing steps
3. Statistical assumptions
4. Pairwise (two-group) comparisons
5. Study questions

STATISTICAL ASSUMPTIONS

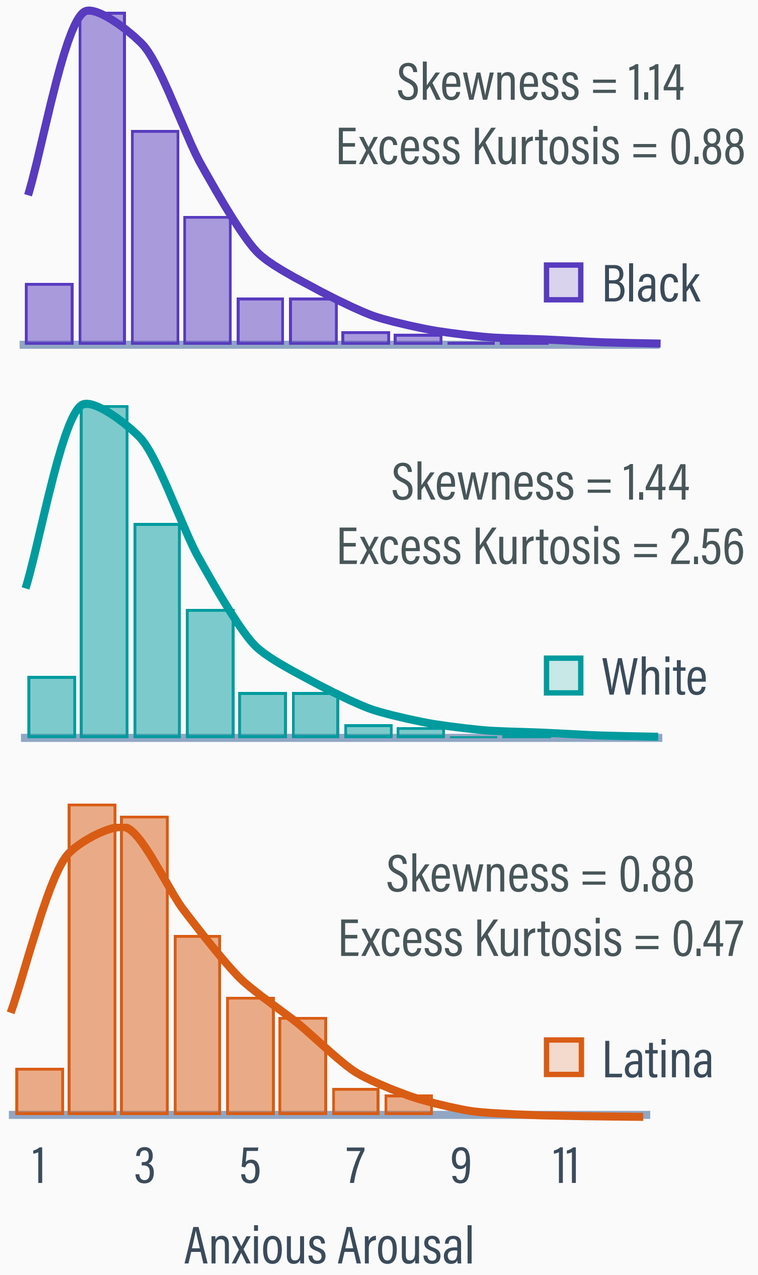
- The accuracy of F-tests (and other statistics) depends on certain conditions in the data being true (e.g., normality)
- Violations of assumptions can bias estimates, inflate or deflate standard errors, and distort significance tests
- Always check reasonableness of assumptions before drawing conclusions

ANOVA ASSUMPTIONS

- Numeric (approximately continuous) dependent variable
- Scores are approximately normal in each population
- Independence of observations (no participant's score influences any other participant's score)
- The populations have equal variances (homogeneity of variance)

WITHIN-GROUP NORMALITY

- In small samples, normality violations can artificially inflate or deflate standard errors, thus distorting significance tests (this N is large!)
- Normality is less of a concern if the sample size is large enough (e.g., Ns > 40 to 50)
- Normalizing transformations (e.g., the natural log of scores) are common in some domains



R OUTPUT

```
 Descriptive statistics by group
Race: Black
                vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
AnxiousArousal     1 901 3.43 1.96      3    3.19 1.48   1  11    10 1.14     0.88 0.07
------------------------------------------------------------------------------------
Race: Latina
                vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
AnxiousArousal     1 373 3.49 1.61      3    3.32 1.48   1  10     9 0.88     0.47 0.08
------------------------------------------------------------------------------------
Race: White
                vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
AnxiousArousal     1 389 3.07 1.56      3    2.85 1.48   1  10     9 1.44     2.56 0.08
```

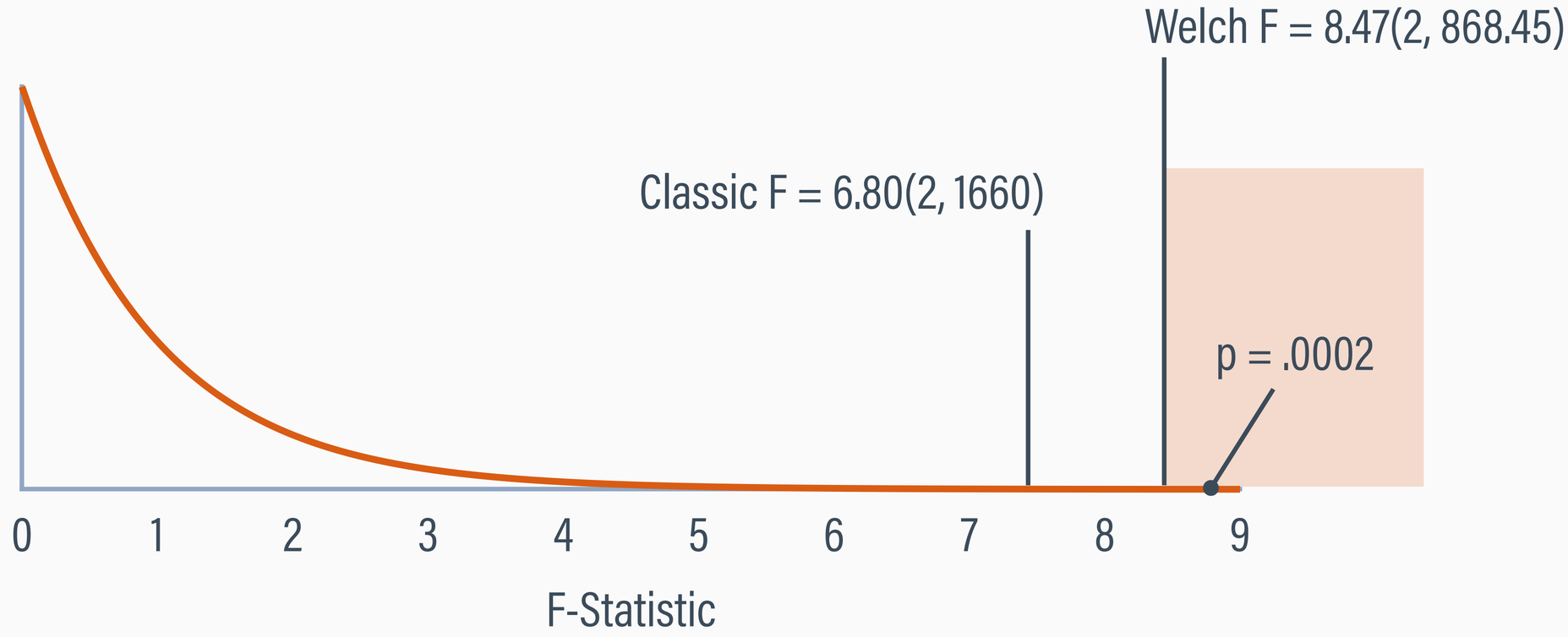
HOMOGENEITY OF VARIANCE

- The classic F-test assumes equal variation; whether and how it affects significance tests depends on the group Ns
- Non-constant variation distorts significance tests (large Ns do not mitigate the problem)
- The Welch F-test we use relaxes this assumption, adjusting both the degrees of freedom and the F-statistic
- Methodological literature shows Welch's test is quite robust to unequal variances and sample sizes

Group	$\bar{X}$	s	n
Black	3.43	1.96	901
White	3.07	1.56	389
Latina	3.49	1.61	373

WELCH F-STATISTIC

- The Welch F-statistic for the PTSD data is noticeably larger because it made a substantial adjustment to the residual df to compensate for the pattern of variances



R OUTPUT

```
    One-way analysis of means (not assuming equal variances)

data:  AnxiousArousal and Race
F = 8.4651, num df = 2.00, denom df = 868.45, p-value = 0.0002286
```

OUTLINE

1. ANOVA overview
2. Significance testing steps
3. Statistical assumptions
4. Pairwise (two-group) comparisons
5. Study questions

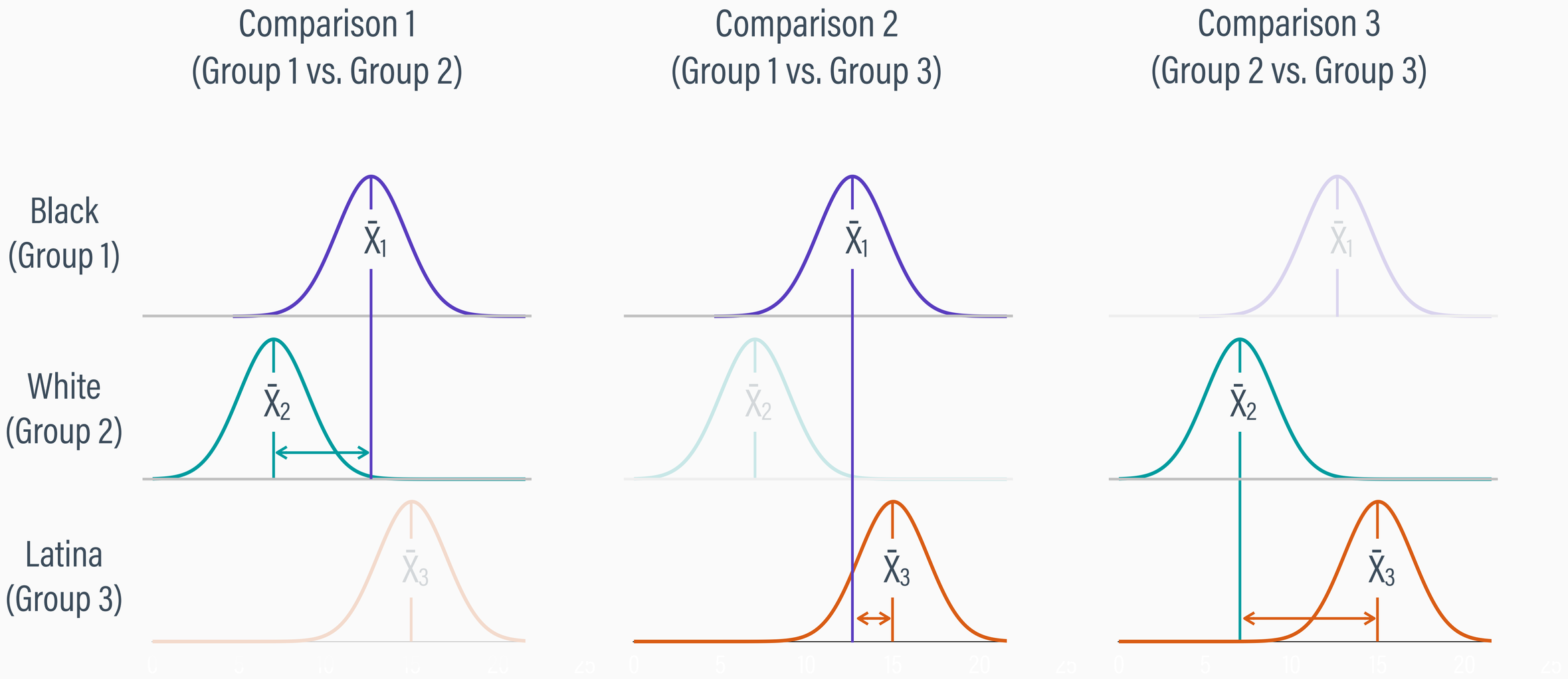
WHAT DOES A SIGNIFICANT F TELL US?

- A significant F test does not tell the whole story about the research question
- Significance implies that at least two groups differ, but it does not tell us which ones
- Any or all pairings of two groups could be different

PAIRWISE OR POST-HOC COMPARISONS

- Researchers usually follow a significant ANOVA by comparing all possible pairs of group means
- These additional analyses are called pairwise (two-group) comparisons or post hoc (“after the fact”) comparisons
- Pairwise comparisons are just multiple independent t-tests

PAIRWISE (TWO-GROUP) COMPARISONS



ACCUMULATING TYPE I ERRORS

- A false positive (a Type I error) occurs when we incorrectly reject the null when it is actually true in the population (e.g., the data lead us to conclude there is an effect when none truly exists)
- If we adopt a 5% rejection criterion for each post hoc test, then the chance of making at least one false positive conclusion accumulates over tests
- With three post hoc tests, there is a roughly 15% chance of at least one false positive (called the familywise Type I error rate)

POST-HOC CORRECTIVE PROCEDURES

- Numerous post-hoc testing procedures have been developed to protect against the accumulation of Type I errors.
- These methods work by inflating the p-value for each test, thereby making it hard to obtain significance and make a Type I error (a false positive can only occur when we reject the null)
- Post-hoc tests differ in how much they inflate p-values and in the degree to which they protect against accumulating Type I errors

TUKEY'S HONESTLY SIGNIFICANT DIFFERENCE

- The Tukey HSD procedure is appropriate when the goal is to perform all possible pairwise (two-group) comparisons
- The HSD t-tests keep the familywise Type I error rate (the cumulative false positive rate) at 5%
- Other procedures are available for scenarios involving other types of comparisons

PAIRWISE COMPARISON RESULTS

- Black and Latina women both have significantly higher PTSD means than White women (both small effect sizes, d > 0.20)
- Black and Latina women do not differ

Group	Mean	SD	n
Black	3.43	1.96	901
White	3.07	1.56	389
Latina	3.49	1.61	373

Group	Mean Diff.	SE Diff.	t	p	Cohen's d
Black vs. White	0.36	0.11	3.35	.002	0.20
Black vs. Latina	−0.06	0.11	−.47	.88	0.03
White vs. Latina	−0.42	0.13	−3.20	.004	0.23

R OUTPUT

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = AnxiousArousal ~ Race, data = PTSD)

$Race
                     diff        lwr        upr     p adj
Latina-Black  0.05240231 -0.2076272  0.3124318 0.8841360
White-Black  -0.36601434 -0.6222339 -0.1097947 0.0023719
White-Latina -0.41841665 -0.7244741 -0.1123592 0.0039054
```

R OUTPUT

```
  .y.             group1 group2 effsize    n1    n2 magnitude
* <chr>           <chr>  <chr>    <dbl> <int> <int> <ord>
1 AnxiousArousal Black  Latina -0.0292   901   373 negligible
2 AnxiousArousal Black  White   0.206    901   389 small
3 AnxiousArousal Latina White   0.264    373   389 small
```

FINAL CONCLUSIONS

- The data lend support to the conclusion that racial disparities in PTSD exist among women at the population level
- Black and Latina women have significantly higher anxious arousal symptoms than White women
- Effect sizes were small, meaning that mean differences are subtle (note that small effect does not mean unimportant!)

APA STYLE WRITE-UP, CONTINUED

We next performed Tukey follow-up tests to examine all possible pairwise group differences. These tests indicated that Black women had significantly higher symptoms than White women ($p = .002$, $d = 0.20$), as did Latina women ($p = .004$, $d = 0.23$.). Finally, Black women and Latina women did not differ in their average symptom levels ($p = .88$, $d = .03$).

OUTLINE

1. ANOVA overview
2. Significance testing steps
3. Statistical assumptions
4. Pairwise (two-group) comparisons
5. Study questions

STUDY QUESTIONS

For these study questions, consider a smoking cessation study that randomly assigns a sample of participants into three conditions: nicotine replacement (e.g., a patch), a call-in quit line support condition, and a “cold turkey” control group condition that receives no assistance. At the end of the study, participants keep a diary of their smoking behavior, and the number of cigarettes smoked during the week following the intervention serves as the outcome (dependent) variable.

STUDY QUESTIONS (1)

1) What are the null and alternate hypotheses for the ANOVA?

2) What is between-group variation measure in the context of this particular study? Describe the sum of squares calculation for this source of variation and how its converted to a mean squares.

3) What is residual variation measure in the context of this particular study? Describe the sum of squares calculation for this source of variation and how its converted to a mean squares.

STUDY QUESTIONS (2)

4) The researchers report the effect size as $R^2 = .10$. Provide an interpretation of the R^2 and describe its magnitude.

5) Suppose that the F-statistic was 9. Provide an interpretation of the F-statistic (not whether it is significant). Explain what the magnitude of the F-statistic indicates about the plausibility of the null hypothesis.

STUDY QUESTIONS (3)

6) The p-value from the F-statistic was p = .01. Provide an interpretation of the probability value (I am not asking whether the test is significant).

7) After finding a significant F-statistic, researchers decide to compare every pair of group means using post-hoc t-tests. Describe the concept of a Type I error in the context. What does it mean to make a Type I error, and how does it happen? Describe the concept of accumulating Type I errors.