MODULE 11

CORRELATION

OUTLINE

1. Correlation overview
2. Significance testing steps
3. Correlation matrices for multiple variables
4. Statistical assumptions
5. Study questions

OUTLINE

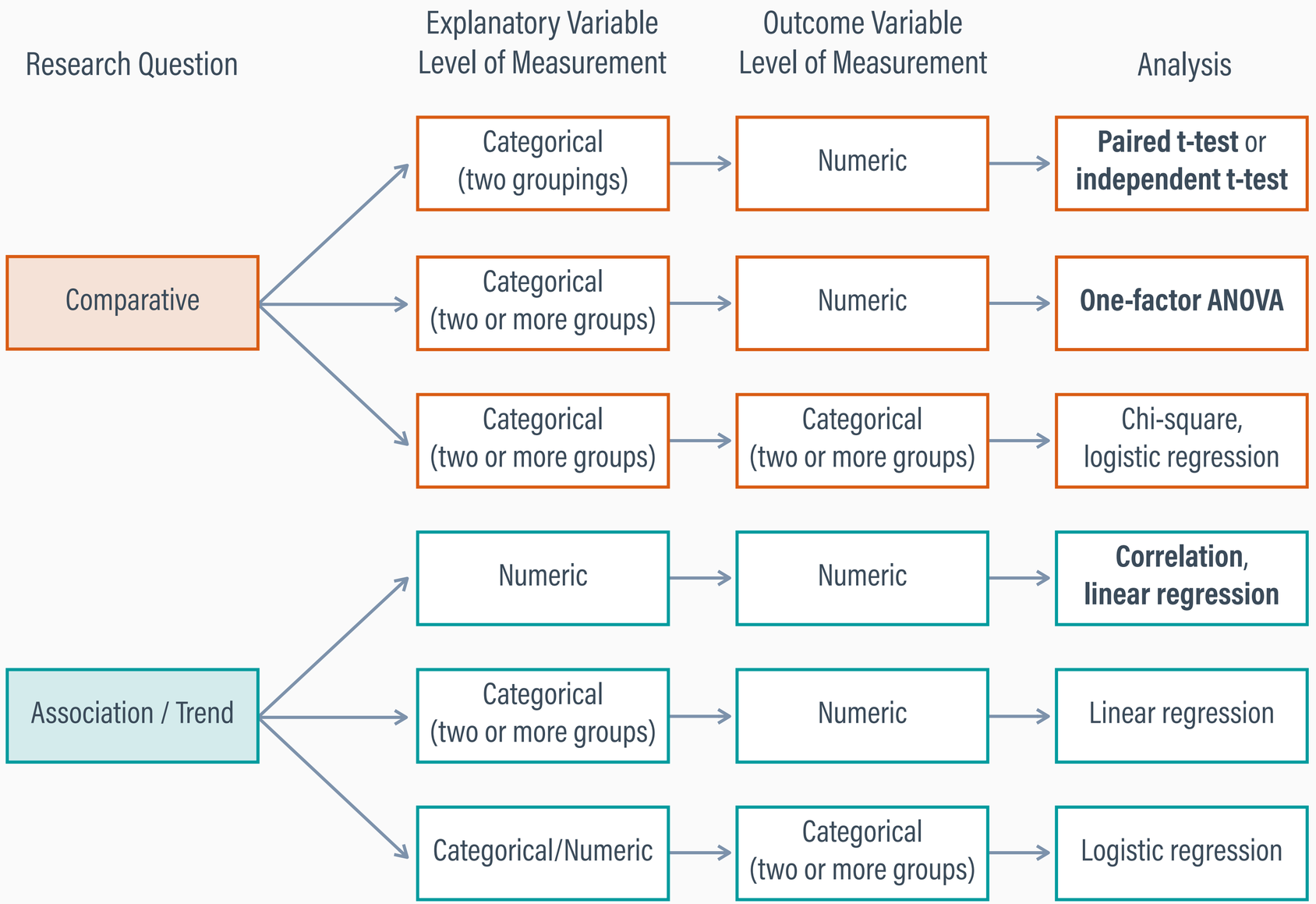
1. Correlation overview
2. Significance testing steps
3. Correlation matrices for multiple variables
4. Statistical assumptions
5. Study questions

CORRELATION

- The colloquial use of “correlation” conveys the idea that two events tend to occur together
- A correlation statistics describes the direction and strength of the association between two variables
- Applicable to association research questions involving a trend between two numeric variables

STATISTICAL ORG CHART

Bold typeface = 250A topic

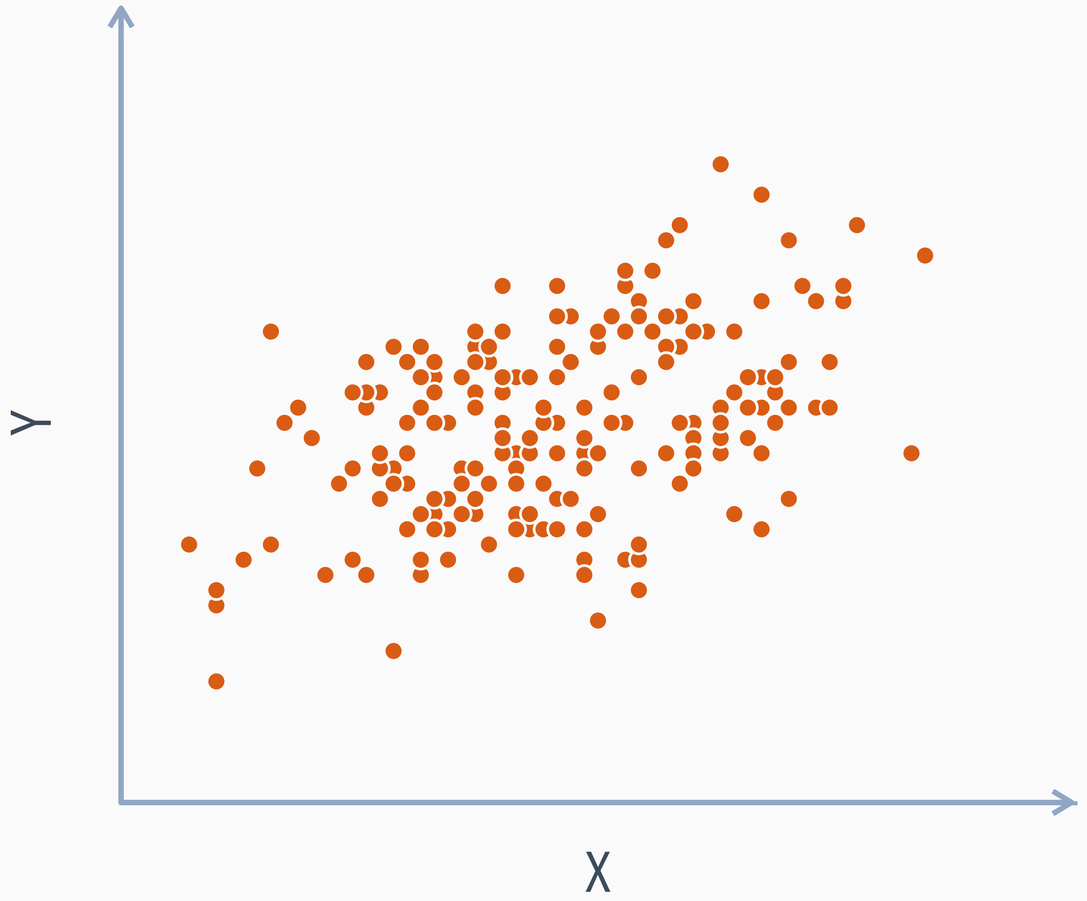


TYPES OF CORRELATIONS

- Pearson's correlation for pairs of numeric variables is the most common correlation in psychology research
- Correlations are available for other metrics: point biserial (binary with numeric), polychoric (ordinal with ordinal), phi (binary with binary)
- We focus on Pearson's correlation

SCATTERPLOTS

- A scatterplot helps visualize a correlation
- Two numeric variables are displayed in a coordinate system, with number lines on the horizontal and vertical axes
- A dot gives the location of each score pair

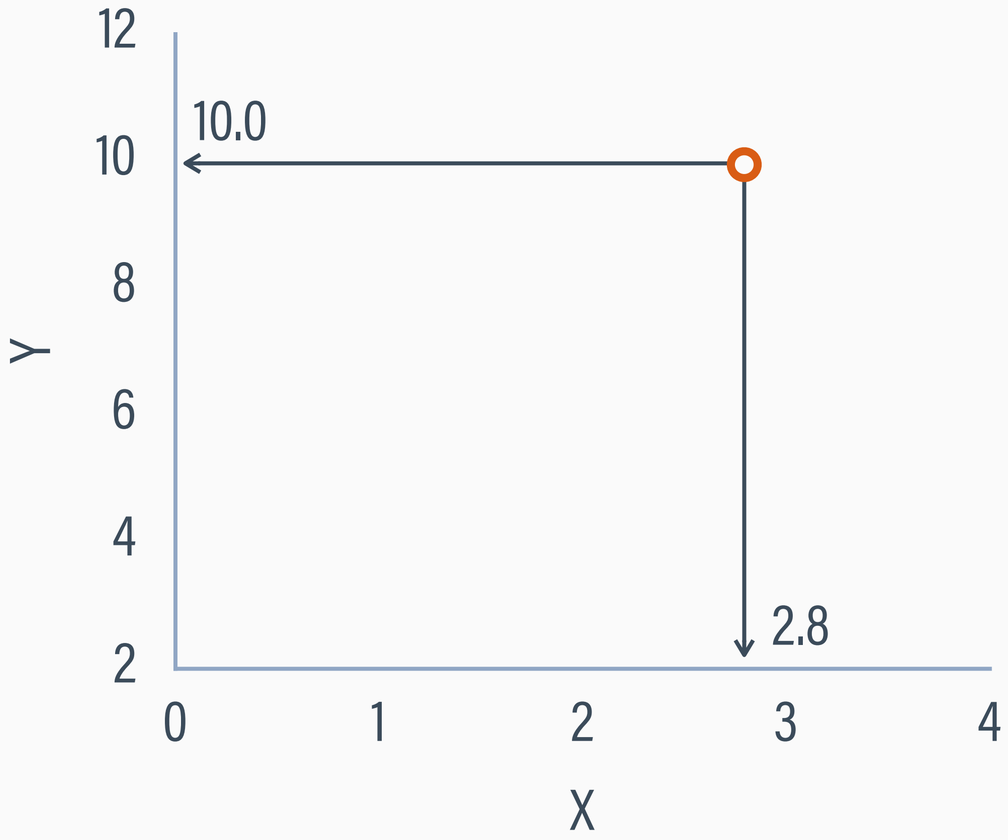


DEMONSTRATION DATA

- Correlation requires two scores per person
- Both variables must be numeric (approximately continuous) because the goal is to describe a linear trend

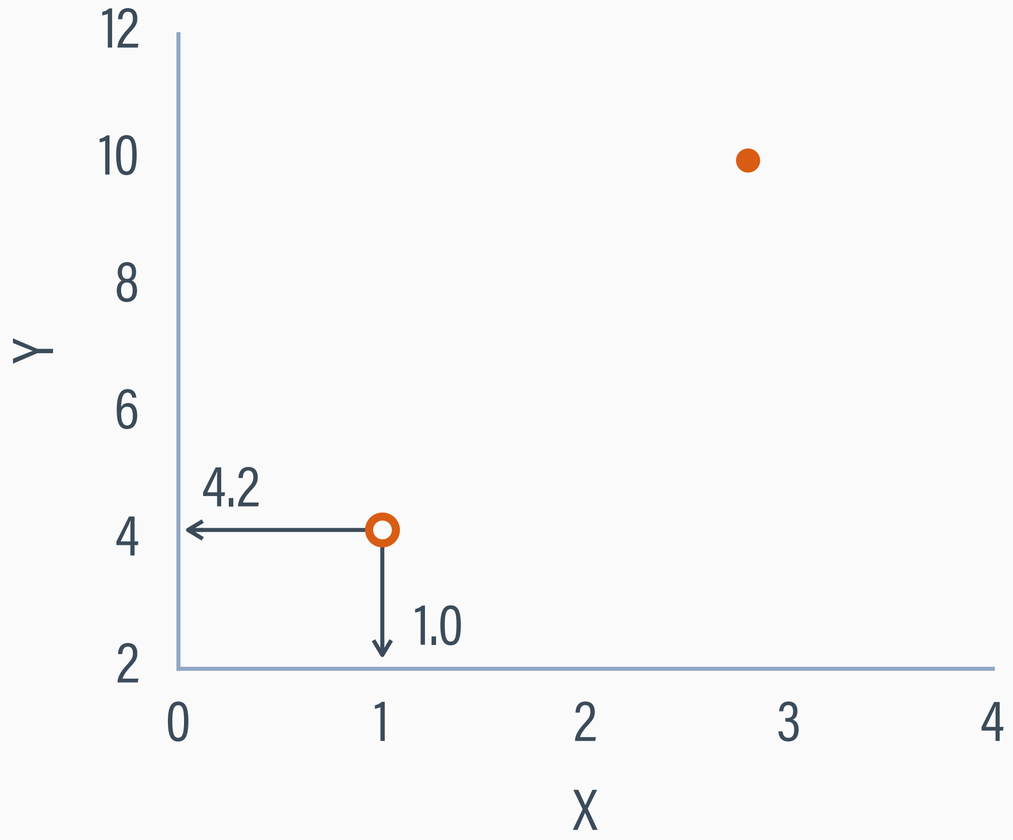
X	Y
2.8	10.0
1.0	4.2
0.1	4.1
3.9	7.0
0.7	5.6
2.6	4.6
0.5	7.5
3.2	6.7
1.8	8.1
1.7	9.5

SCORE PAIR 1



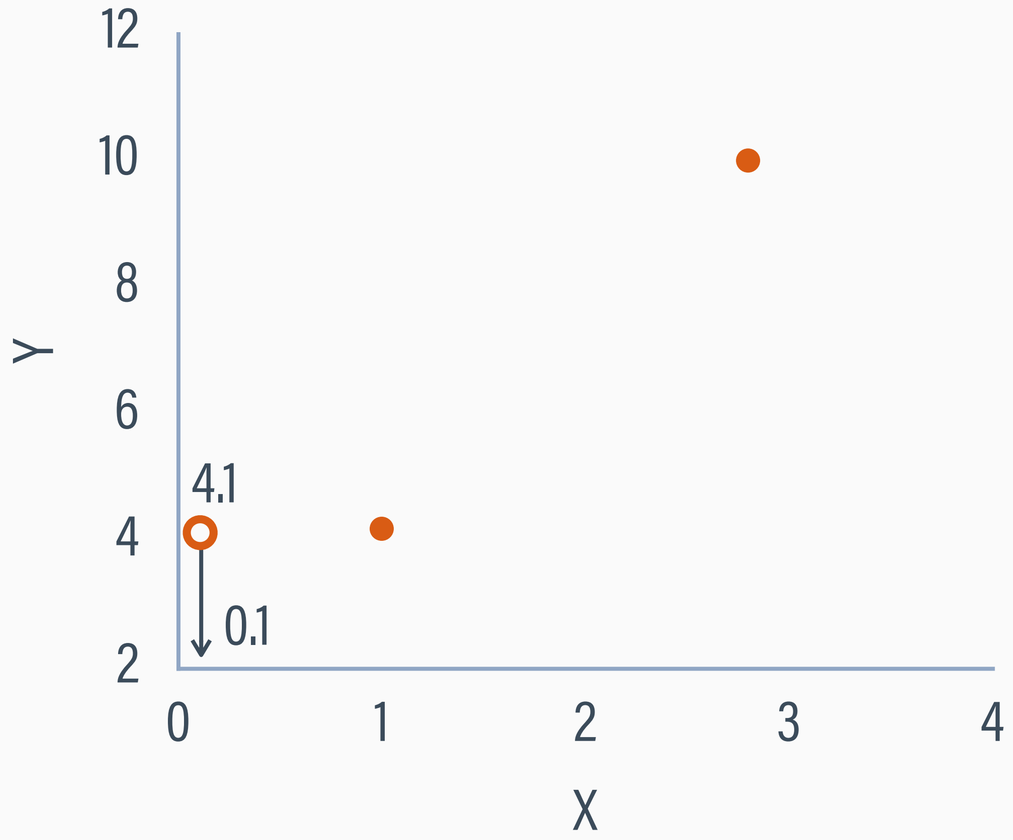
X	Y
2.8	10.0
1.0	4.2
0.1	4.1
3.9	7.0
0.7	5.6
2.6	4.6
0.5	7.5
3.2	6.7
1.8	8.1
1.7	9.5

SCORE PAIR 2



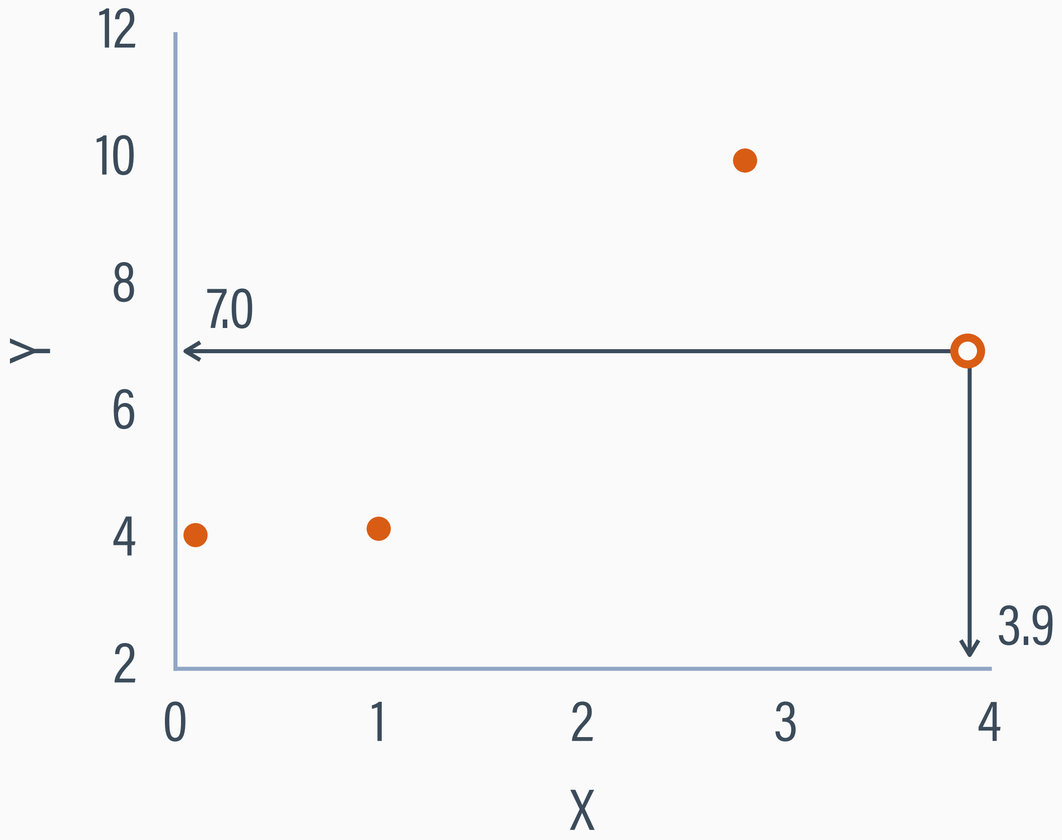
X	Y
2.8	10.0
1.0	4.2
0.1	4.1
3.9	7.0
0.7	5.6
2.6	4.6
0.5	7.5
3.2	6.7
1.8	8.1
1.7	9.5

SCORE PAIR 3



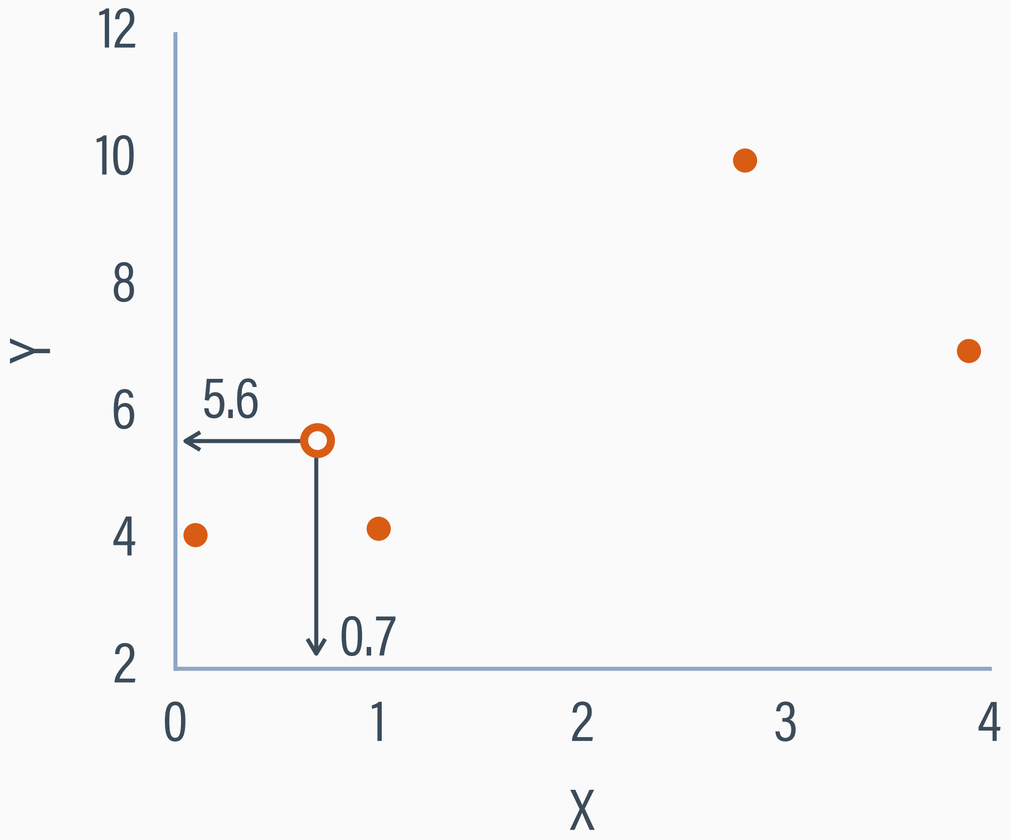
X	Y
2.8	10.0
1.0	4.2
0.1	4.1
3.9	7.0
0.7	5.6
2.6	4.6
0.5	7.5
3.2	6.7
1.8	8.1
1.7	9.5

SCORE PAIR 4



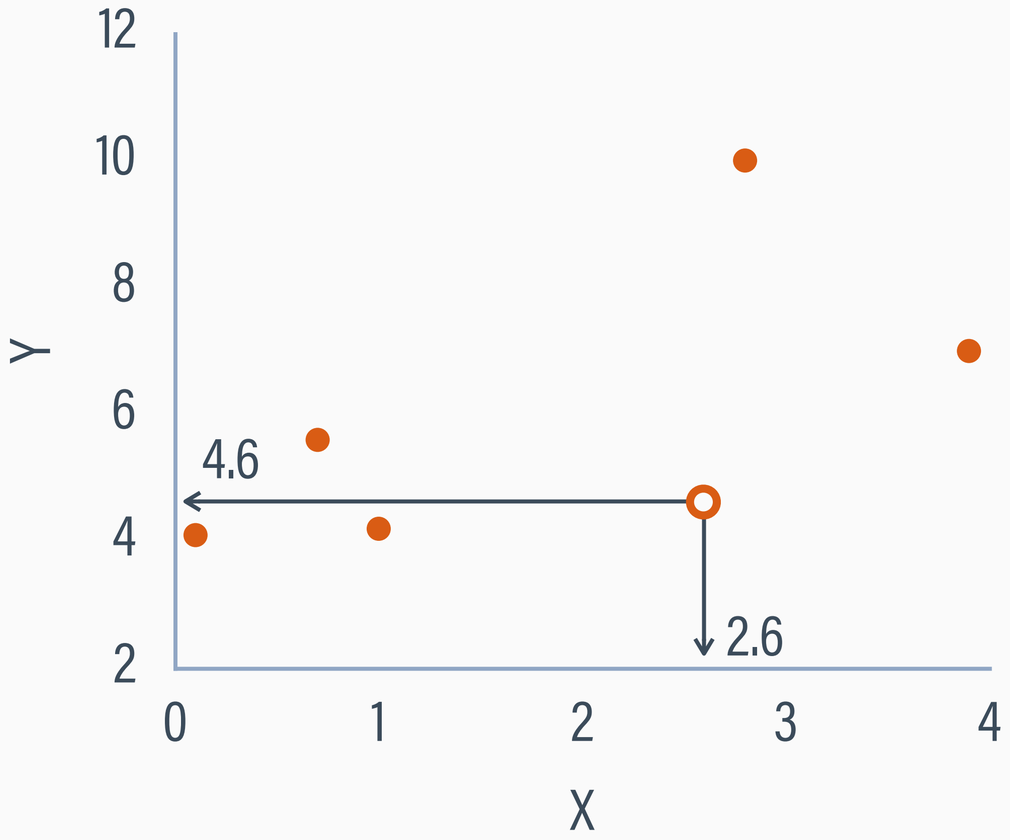
X	Y
2.8	10.0
1.0	4.2
0.1	4.1
3.9	7.0
0.7	5.6
2.6	4.6
0.5	7.5
3.2	6.7
1.8	8.1
1.7	9.5

SCORE PAIR 5



X	Y
2.8	10.0
1.0	4.2
0.1	4.1
3.9	7.0
0.7	5.6
2.6	4.6
0.5	7.5
3.2	6.7
1.8	8.1
1.7	9.5

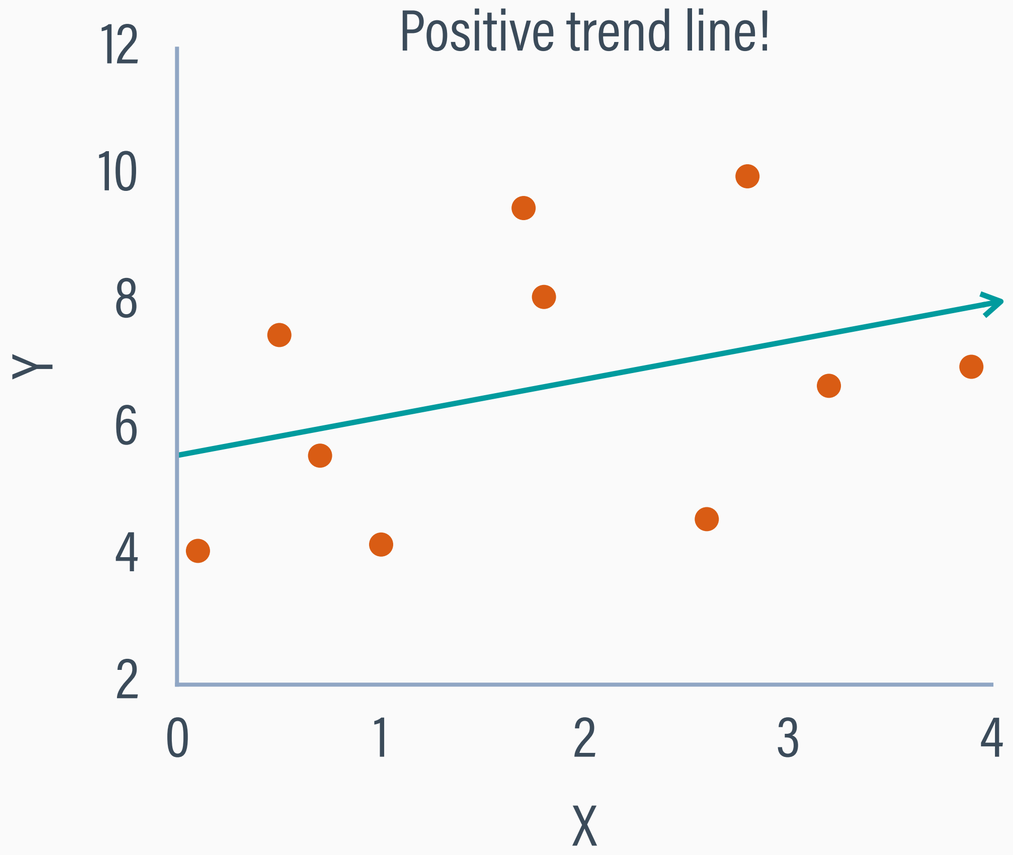
SCORE PAIR 6



X	Y
2.8	10.0
1.0	4.2
0.1	4.1
3.9	7.0
0.7	5.6
2.6	4.6
0.5	7.5
3.2	6.7
1.8	8.1
1.7	9.5

FINAL SCATTERPLOT

- There is a positive trend: Y increases as X increases, Y decreases as X decreases
- A straight line describes an overall positive trend, but the data do not fall on the line
- A correlation indexes the trend's direction and the closeness of the data to a straight line



PEARSON'S CORRELATION COEFFICIENT

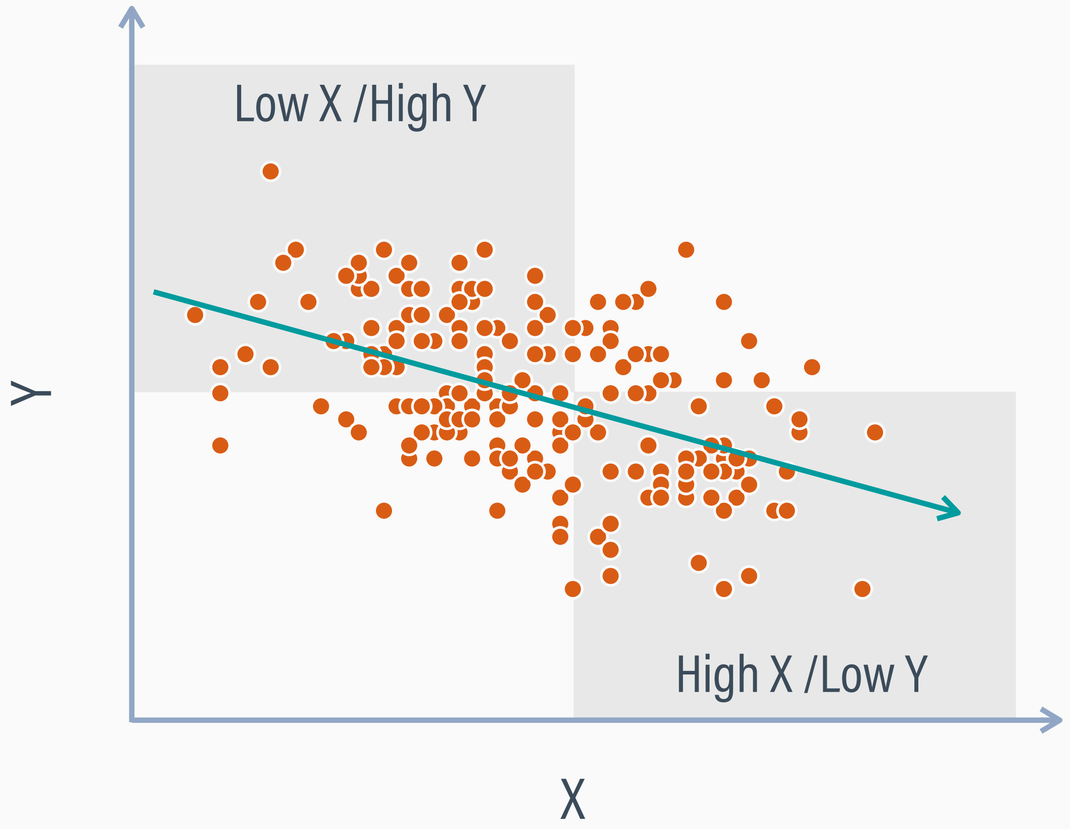
- Pearson's correlation quantifies the linear association between two numeric variables (the degree to which the data approximate a straight line)
- The direction of the correlation can be positive or negative, and its strength ranges from 0 (no association or trend) to ±1 (data fall exactly on a line)
- ρ = population correlation, r = sample correlation

POSITIVE CORRELATION

- Scores move in the same direction
- A high score on one variable tends to pair with a high score on the other variable
- A low score on one variable tends to pair with a low score on the other variable

NEGATIVE CORRELATION

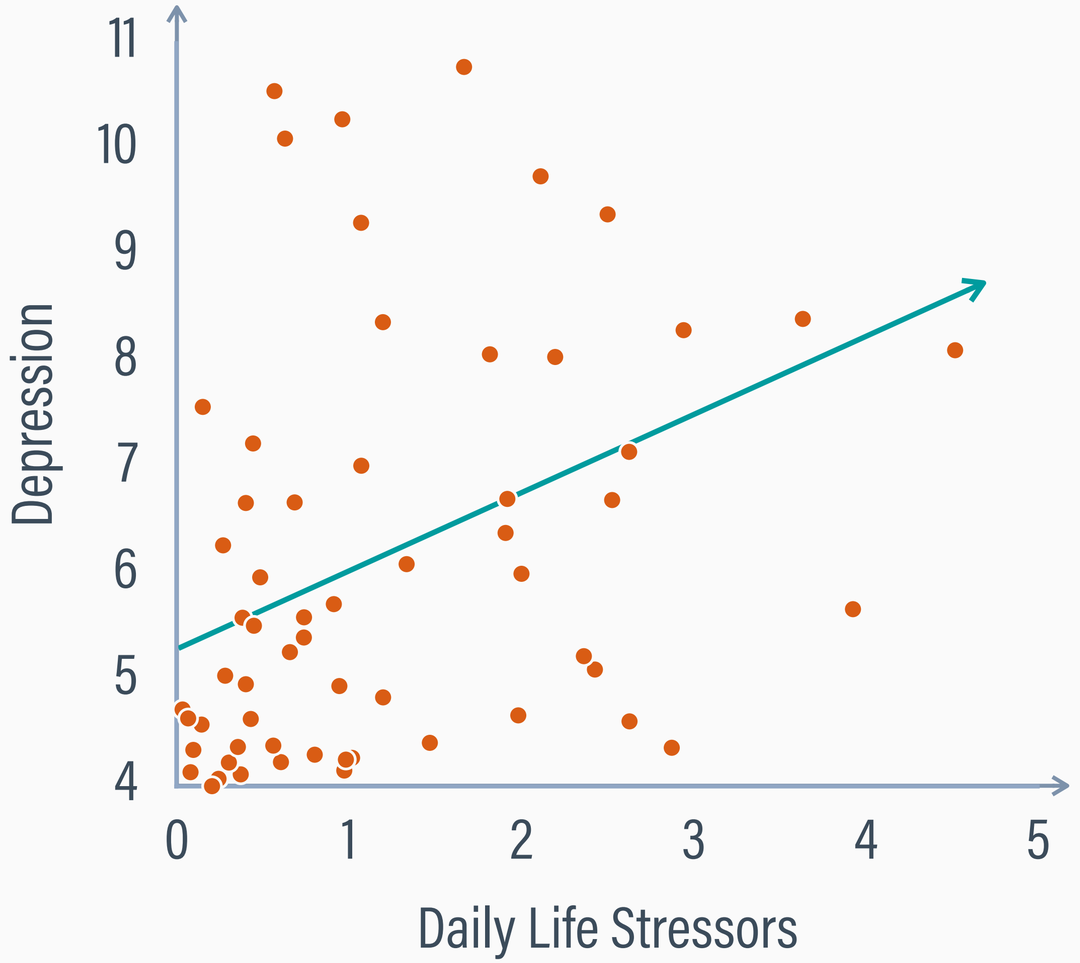
- Scores move in the opposite direction
- A high score on the first variable tends to pair with a low score on the second variable
- A low score on the first variable tends to pair with a high score on the second



In small groups of two or three, generate pairs of variables in your own research domains that would likely exhibit a positive correlation. Generate pairs of variables that would likely exhibit a negative correlation.

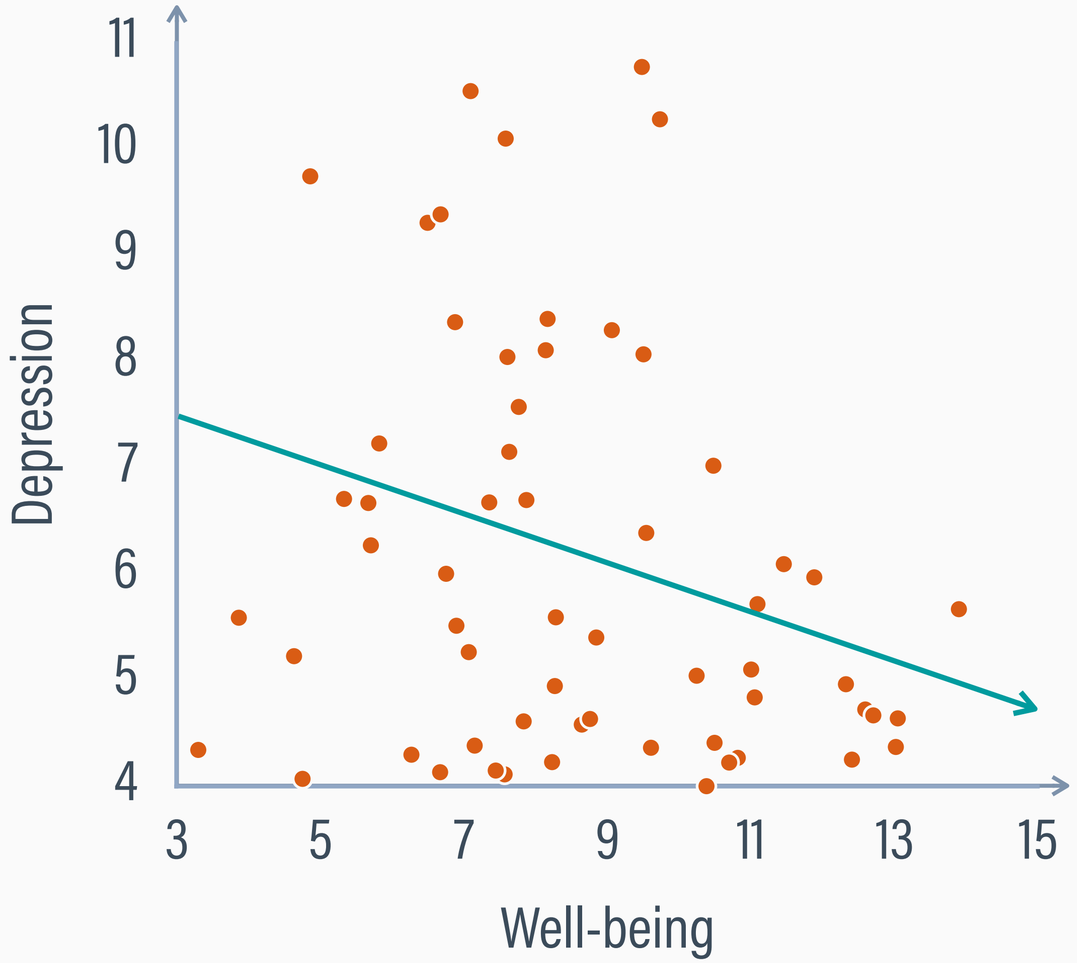
POSITIVE CORRELATION EXAMPLE

- People with more daily stressors tend to have higher depression, and people with fewer stressors tend to have lower depression
- A positively sloped line describes the direction of the overall trend



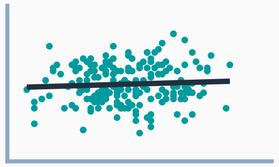
NEGATIVE CORRELATION EXAMPLE

- Researchers also collected daily measures of psychological well-being
- People with higher well-being tend to have lower depression, and people with lower well-being tend to report higher depression
- A negatively sloped line describes the direction of the overall trend

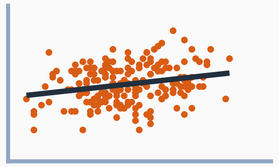


CORRELATION EFFECT SIZE BENCHMARKS

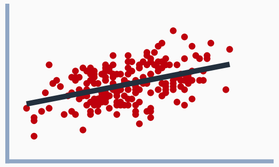
- Pearson's correlation is also a type of effect size



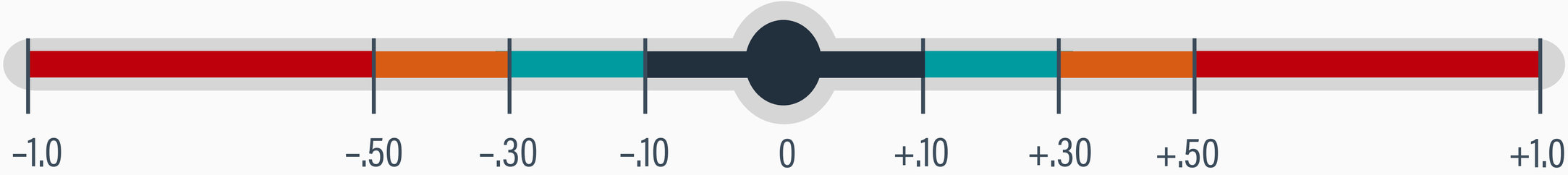
Small
|.10 to .30|



Moderate
|.30 to .50|

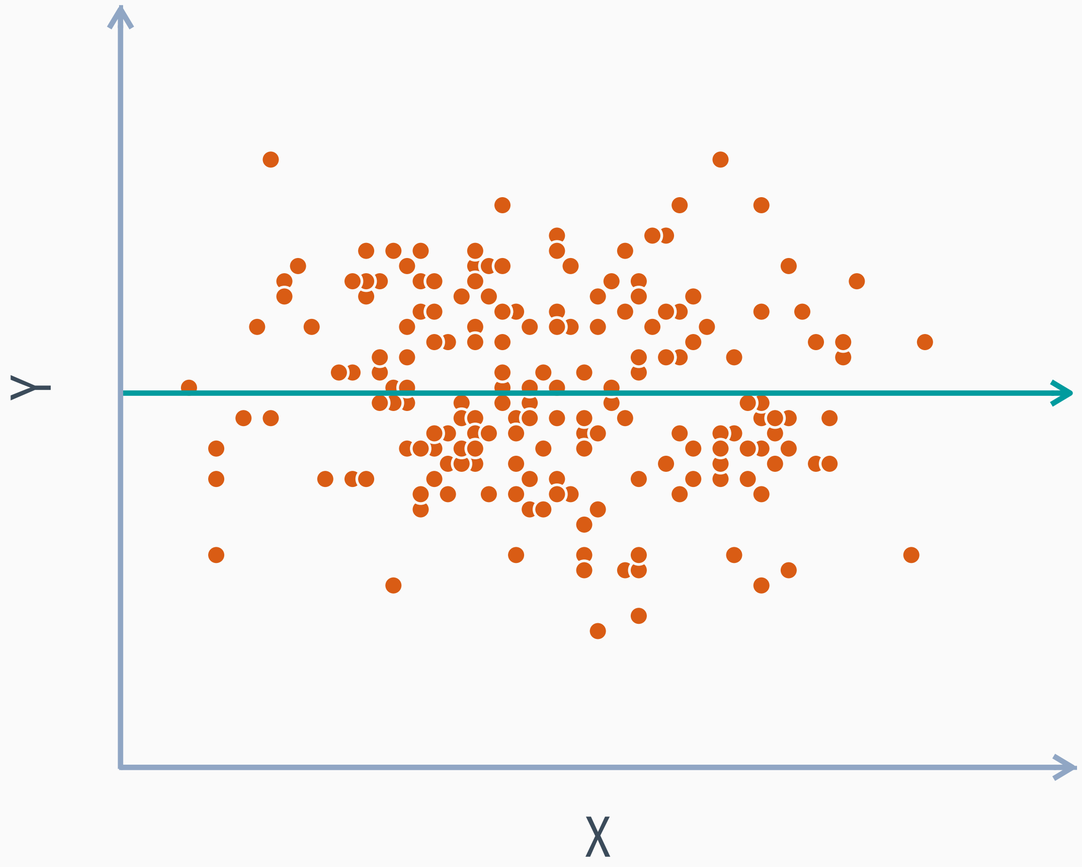


Moderate
|.50 to 1.0|



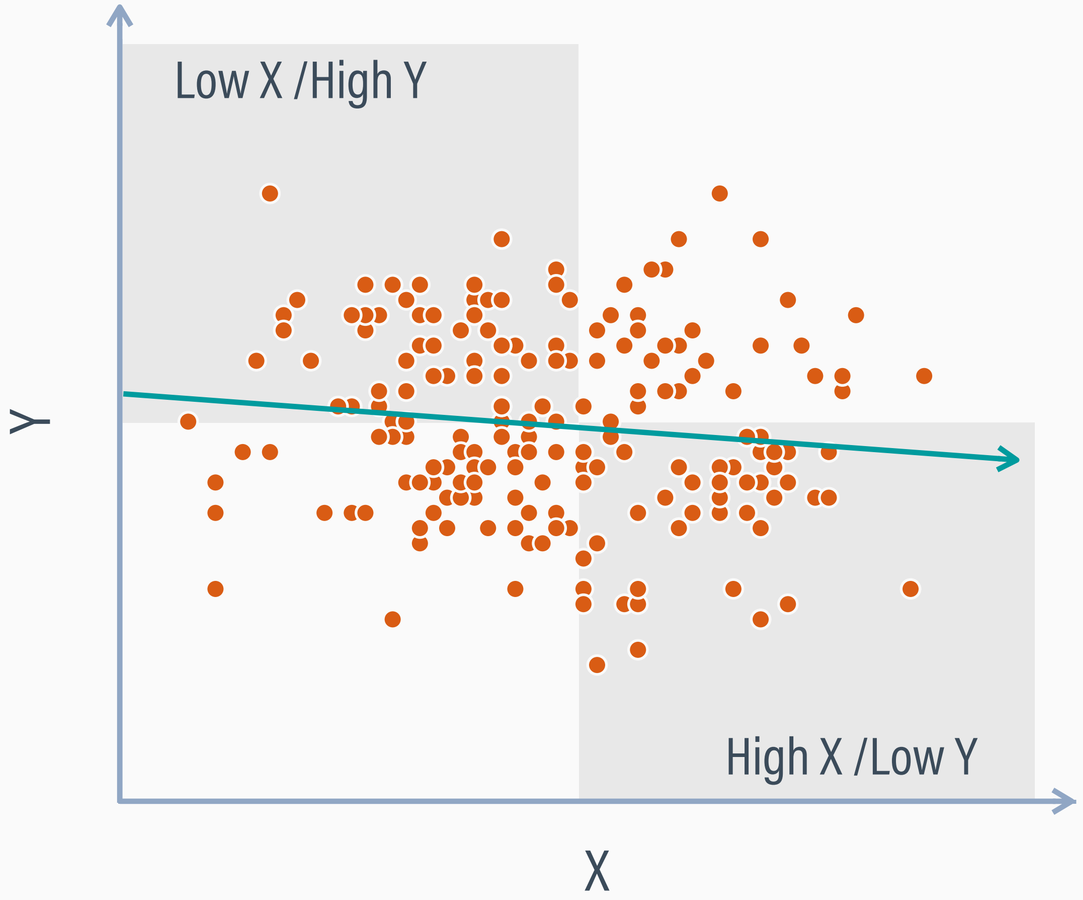
NO CORRELATION: r = 0

- A zero correlation occurs when there is no discernible trend line in the data
- Knowing whether somebody is high or low on X tells us nothing about their Y
- The data are dispersed around a flat line with a slope equal to zero

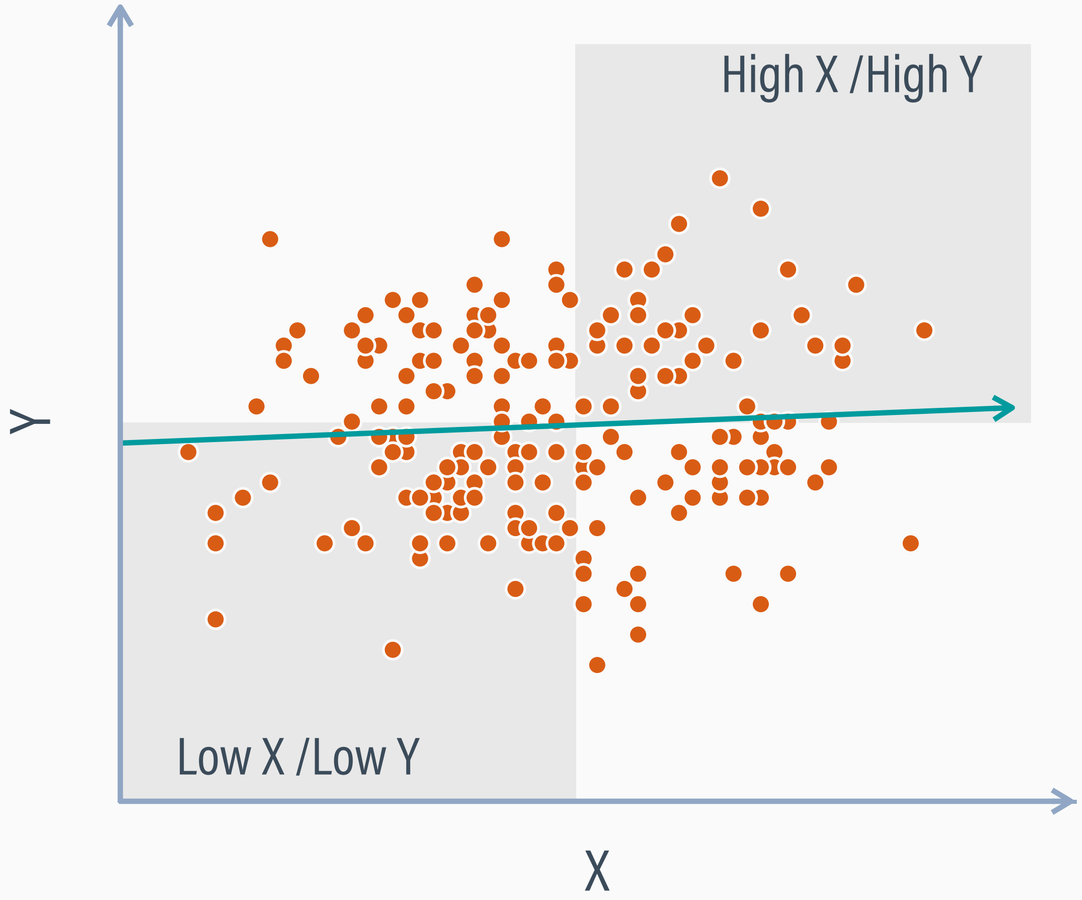


WEAK CORRELATION: r = ± .10

Weak negative (r = –.10)

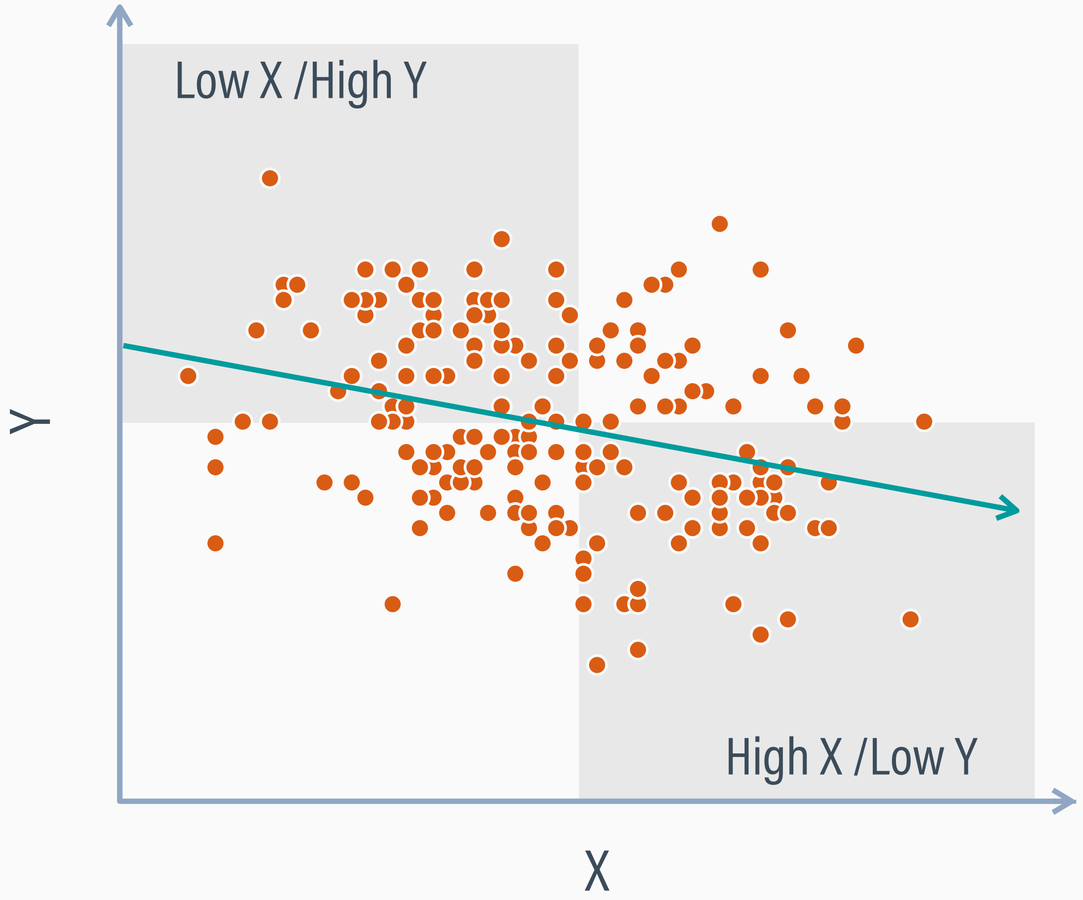


Weak positive (r = .10)

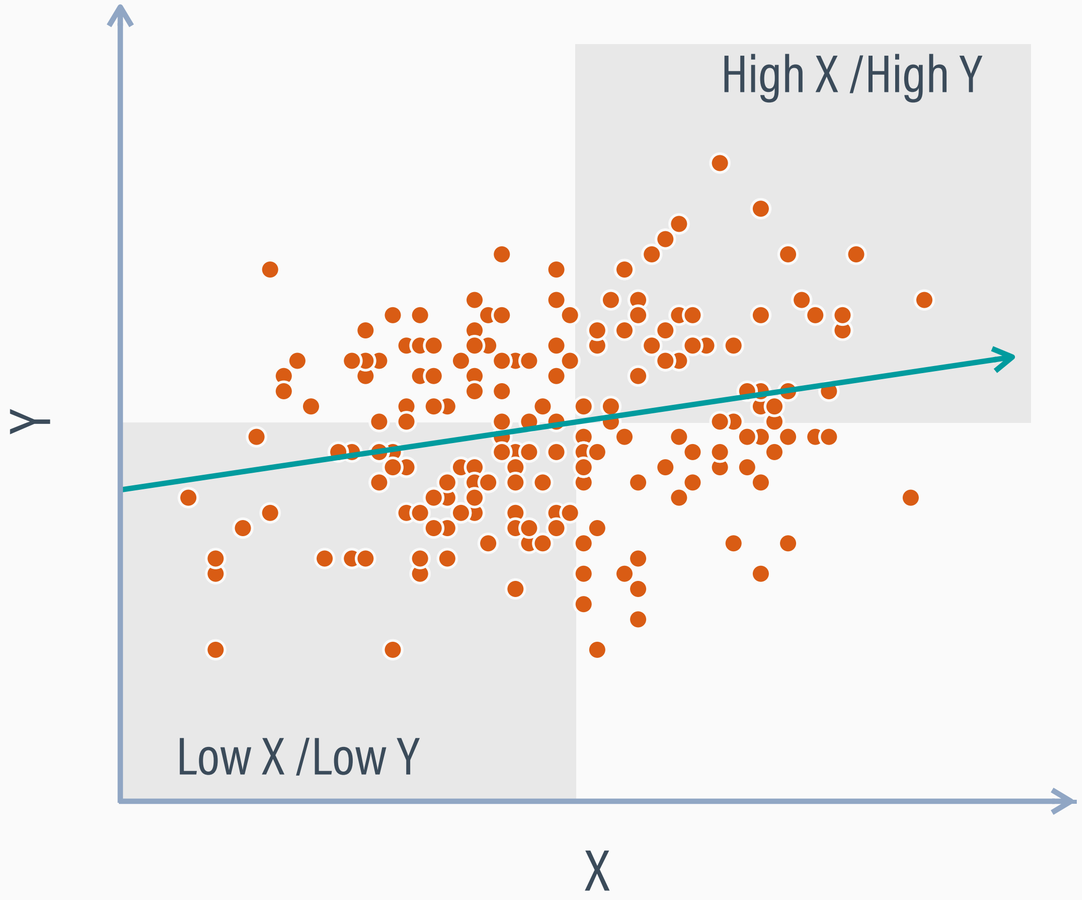


MODERATE CORRELATION: r = ± .30

Moderate negative (r = –.30)



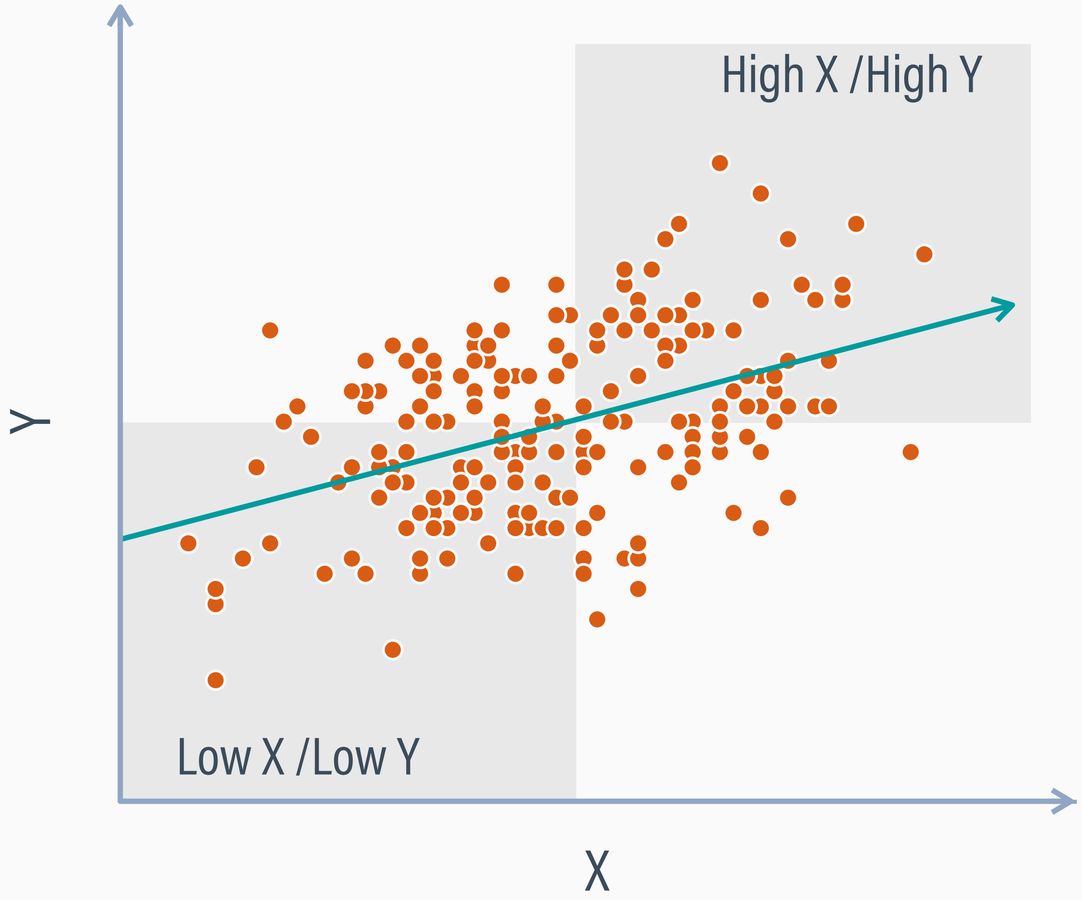
Moderate positive (r = .30)



STRONG CORRELATION: r = ± .50

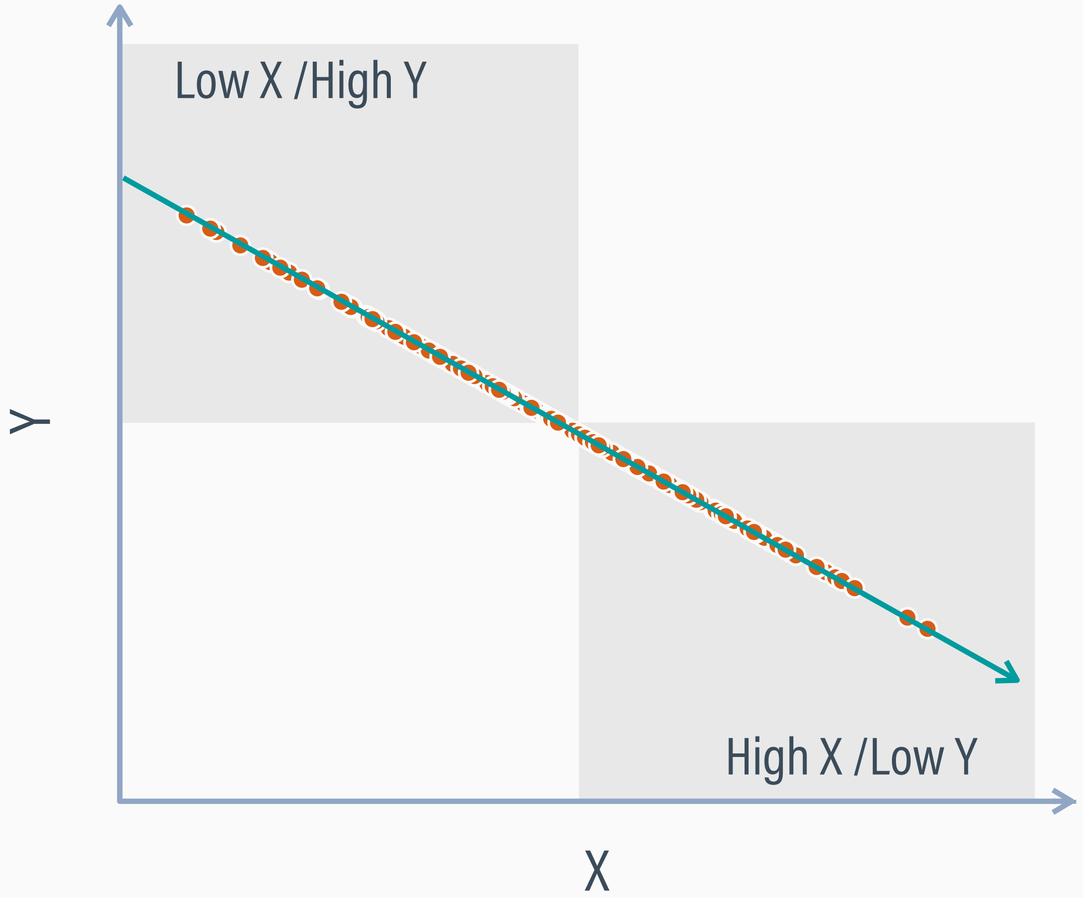
Strong negative (r = –.50)

Strong positive (r = .50)

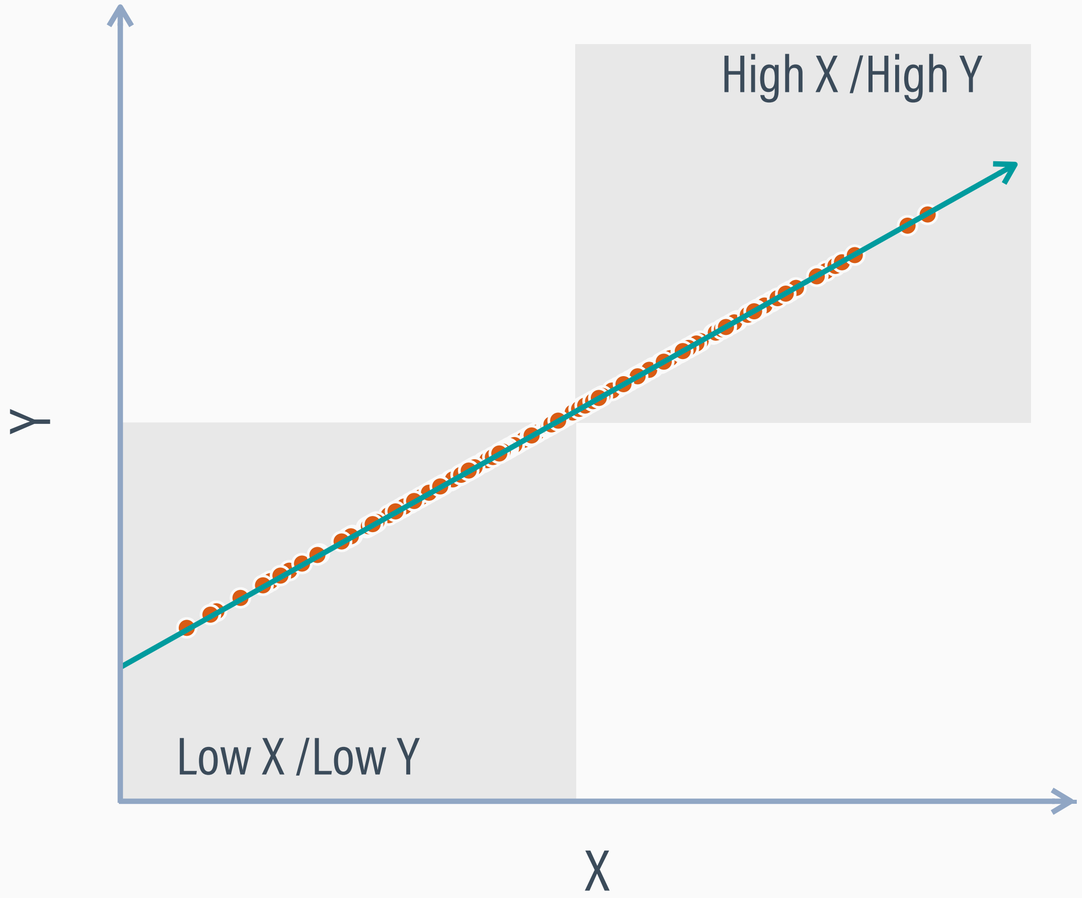


PERFECT CORRELATION: r = ± 1.0

Perfect negative (r = –1.00)



Perfect positive (r = 1.00)



OUTLINE

1. Correlation overview
2. Significance testing steps
3. Correlation matrices for multiple variables
4. Statistical assumptions
5. Study questions

DAILY DIARY STUDY OF MOOD

This study delivered a daily diary to 61 HIV infected men who have sex with men (MSM) between 16 and 24 years old for 66 days to measure HIV-risk behaviors and other psychosocial variables. The study examined the association between daily life stressors and daily negative mood. The study also examined the person-level association between the average number of daily life stressors across the 66 days and the average level of negative mood across that period.

Manuel Ramirez

Patrick Wilson

Ramirez, M, Wilson, P., Mitchell, R., Enders, C., & Woller, M. (in progress). Daily variability in depressed mood among gay and bisexual youth living with HIV. *Manuscript in preparation*.

KEY VARIABLES

Daily Life Stressors

Respondents were presented with list of stressful events (e.g., fights with family or friends, work stress, financial stress), and they checked how many they experienced each day.

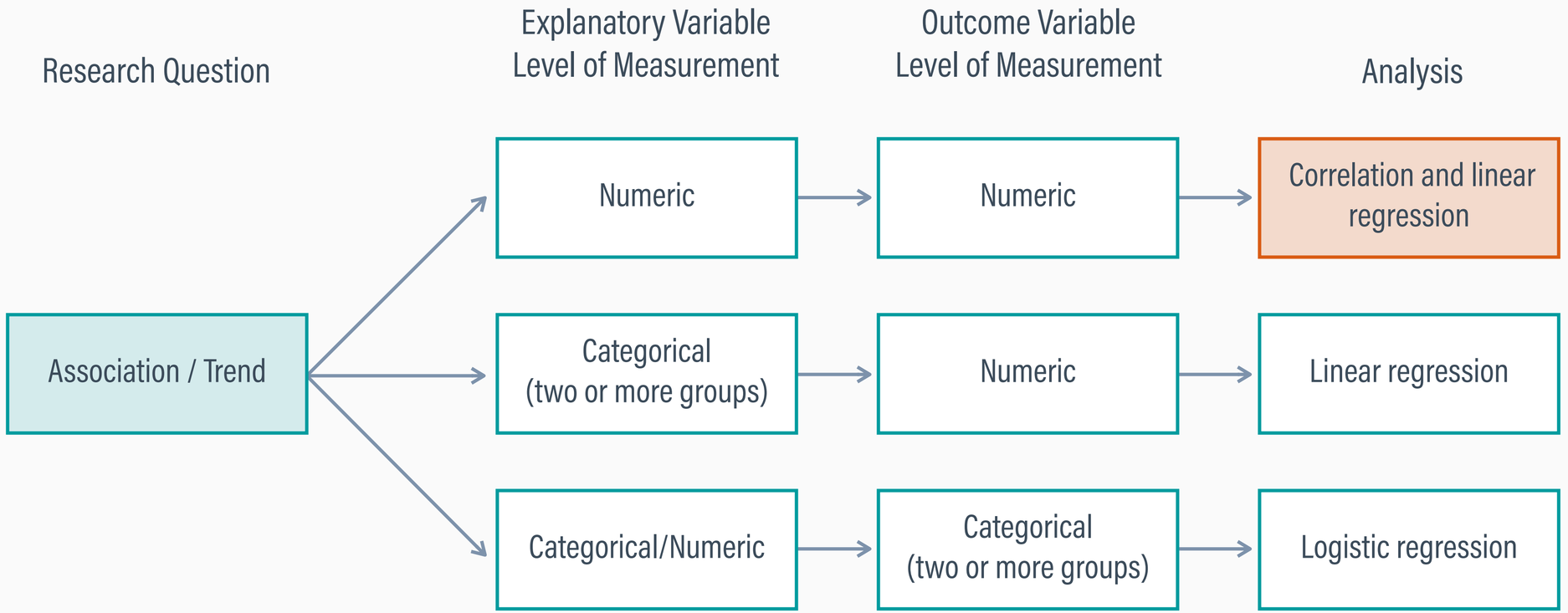
Depression

Depression was measured using the Profile of Mood States (POMS). The POMS Depression-Dejection scale is computed by summing responses to five questionnaire items, where higher ratings reflect more negative mood.

RESEARCH QUESTION

- Question: Is there an association or trend between one's number of life stressors and their average depressive mood?
- The explanatory (independent) variable, number of life stressors, is a numeric value derived from a checklist
- The outcome (dependent) variable, depressive mood, is a numeric scale derived from several questionnaire items

STATISTICAL ORG CHART



SIGNIFICANCE TESTING STEPS

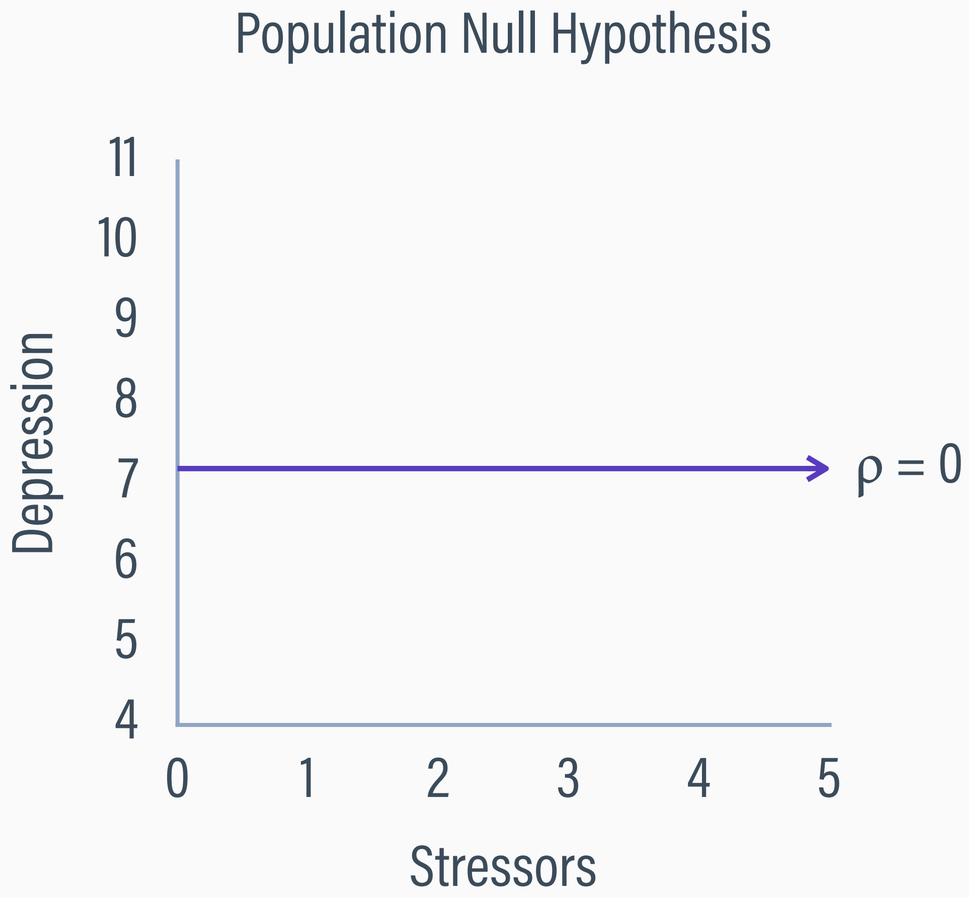
1. Specify hypotheses
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

NULL HYPOTHESIS

- Null hypothesis: In the population, there is no relation between stressors and depression

$$H_0: \rho = 0$$

- The null is counter to expectations because researchers anticipate that changes in stressors could correspond with changes in depression



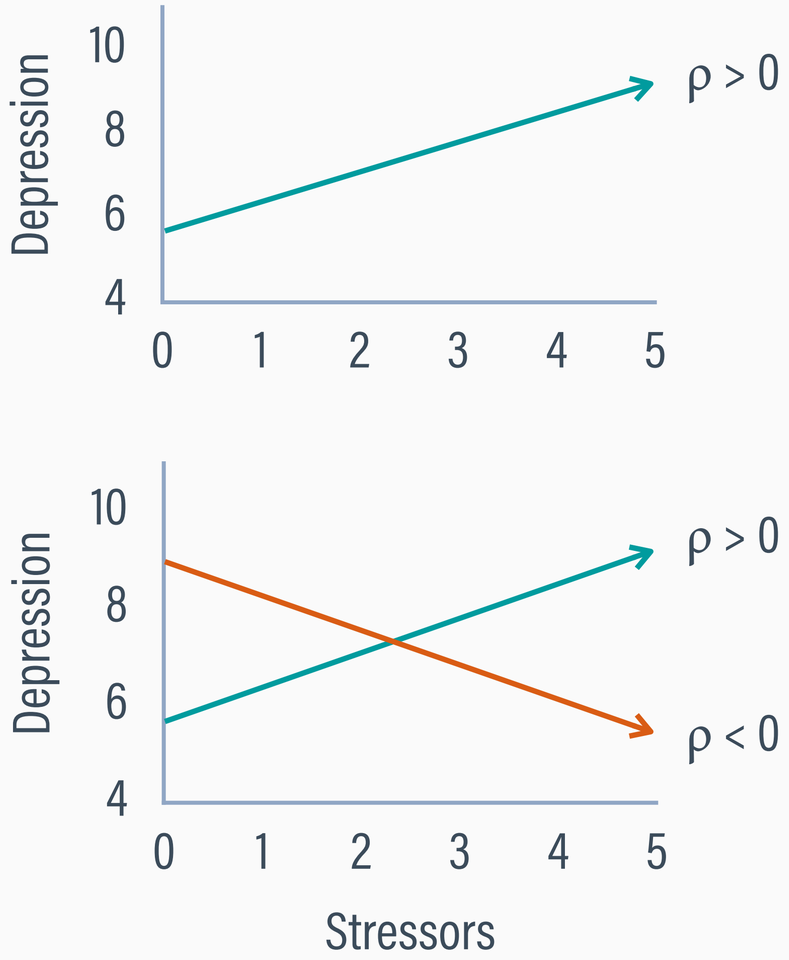
TWO POSSIBLE ALTERNATIVE HYPOTHESES

- One-tailed alternate: An increase in stressors could only be related to a corresponding increase in depression

$$H_A: \rho > 0$$

- Two-tailed alternate: the relation between stressors and depression could be positive or negative

$$H_A: \rho \neq 0 \ (\rho > 0 \text{ or } \rho < 0)$$

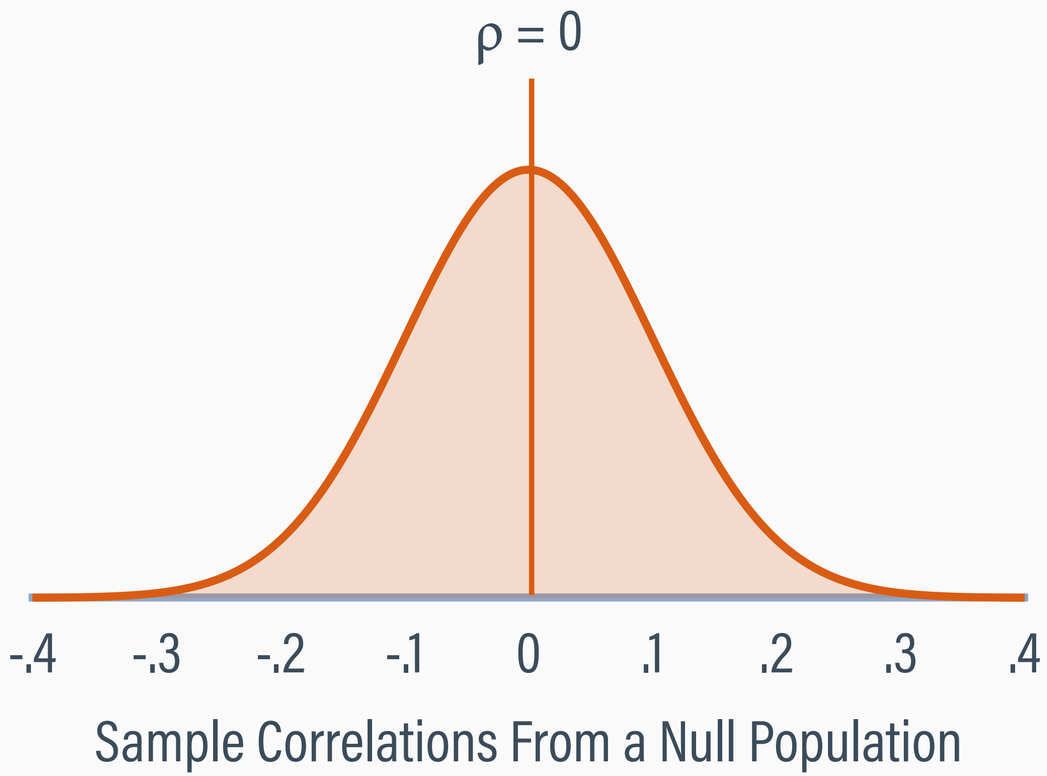


SIGNIFICANCE TESTING STEPS

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

SAMPLING DISTRIBUTION IF THE NULL IS TRUE

- Like means and mean difference statistics, correlations vary across other hypothetical data sets that we could have worked with
- Samples from a null population would produce correlations that are distributed around a true slope of zero

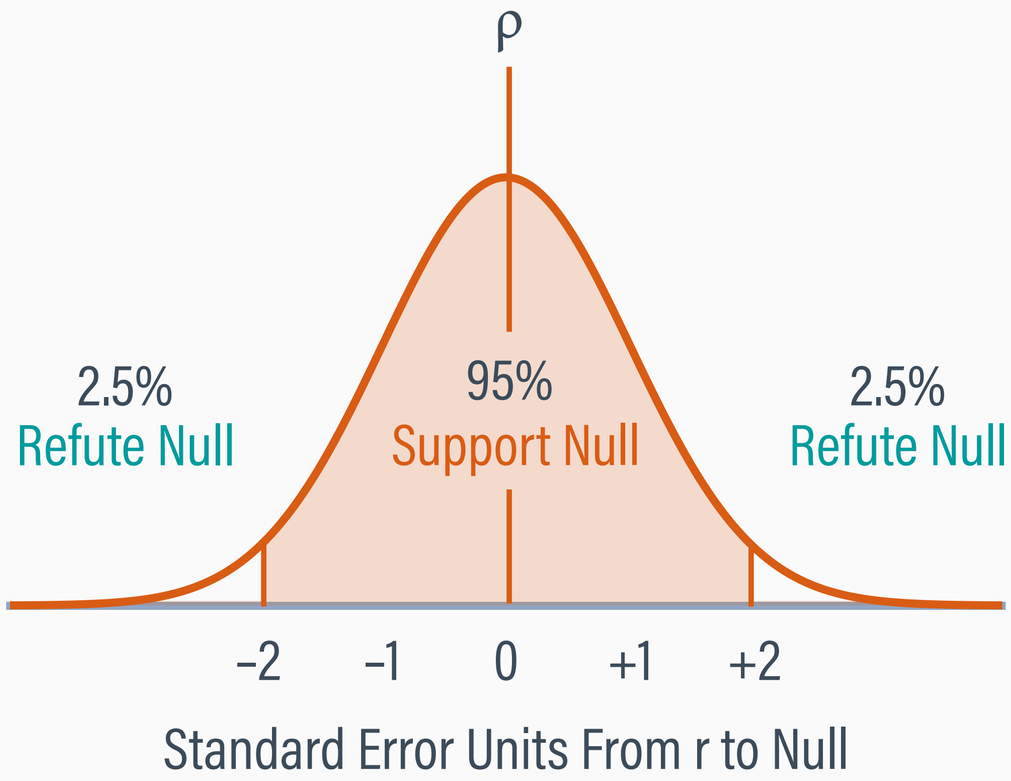


STANDARD OF EVIDENCE

- The data are the evidence that we use to conclude whether the null is plausible (“innocent”) or implausible (“guilty”)
- If the sample correlation from our data (denoted r with Roman letters) is very different from the null correlation ($\rho = 0$), then we conclude that the null hypothesis is implausible
- How big of a Pearson’s r do we need to see to refute the null?

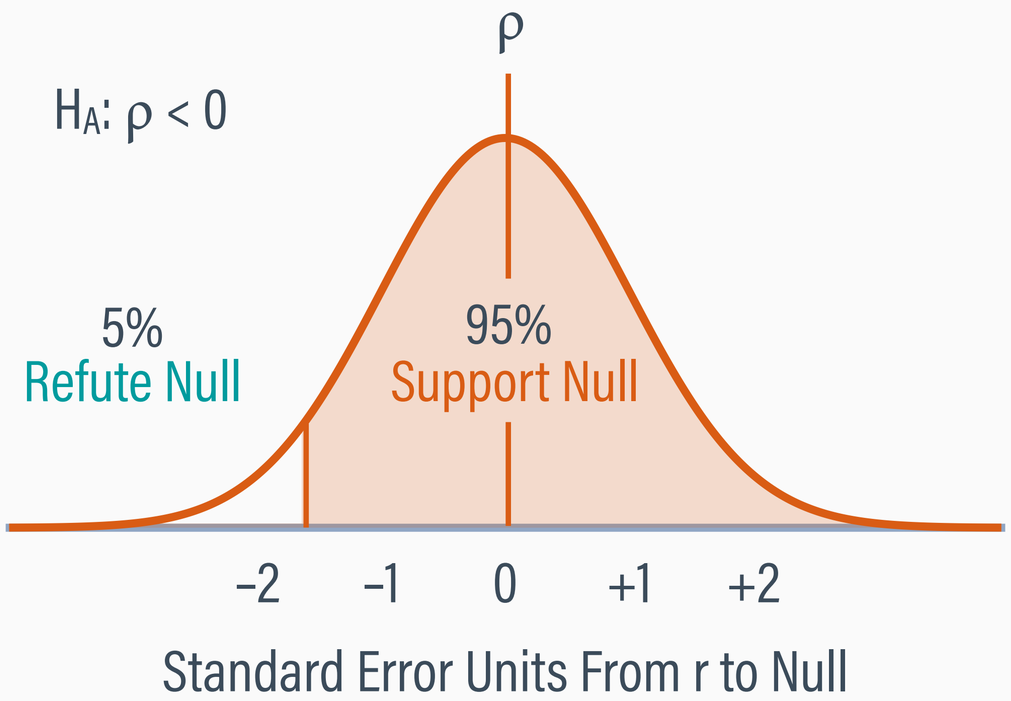
TWO-TAILED ALTERNATE HYPOTHESES

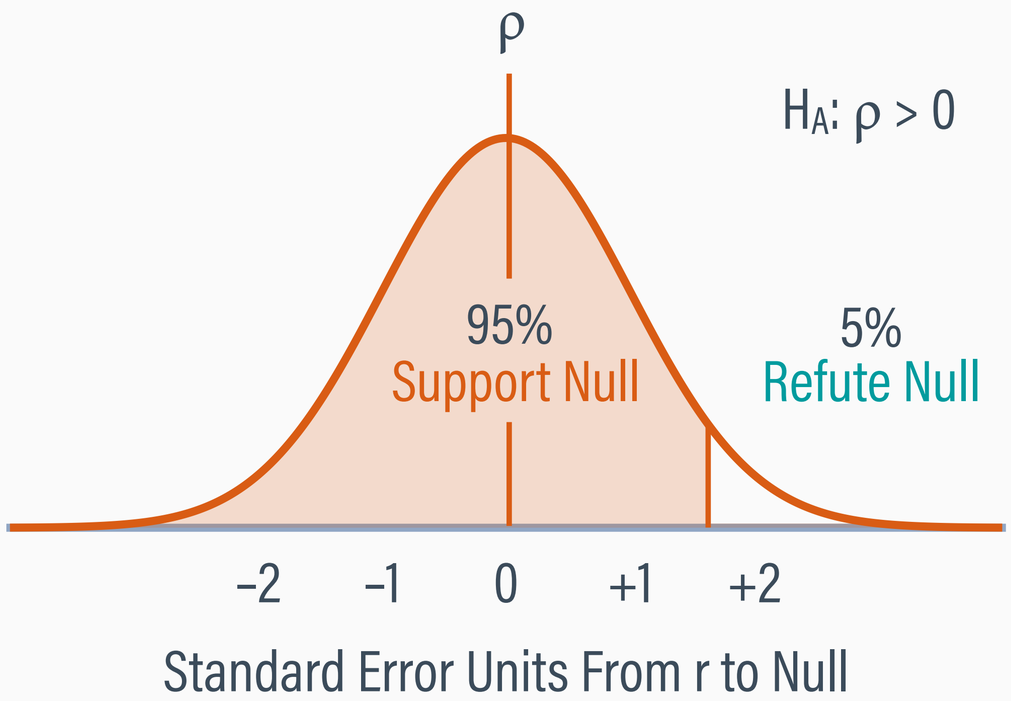
- By convention, we refute the null if the sample correlation r falls outside the middle 95% of the sampling distribution
- Such as sample has less than a 5% chance of originating from the null population ($p < .05$)
- The 5% rejection region (alpha level) is split in half to allow for the possibility that either an increase or a decrease provides evidence against H_0



ONE-TAILED ALTERNATE HYPOTHESES

- The 5% rejection region (alpha level) is placed in one tail, since only a positive (or only a negative) r counts as evidence against H_0



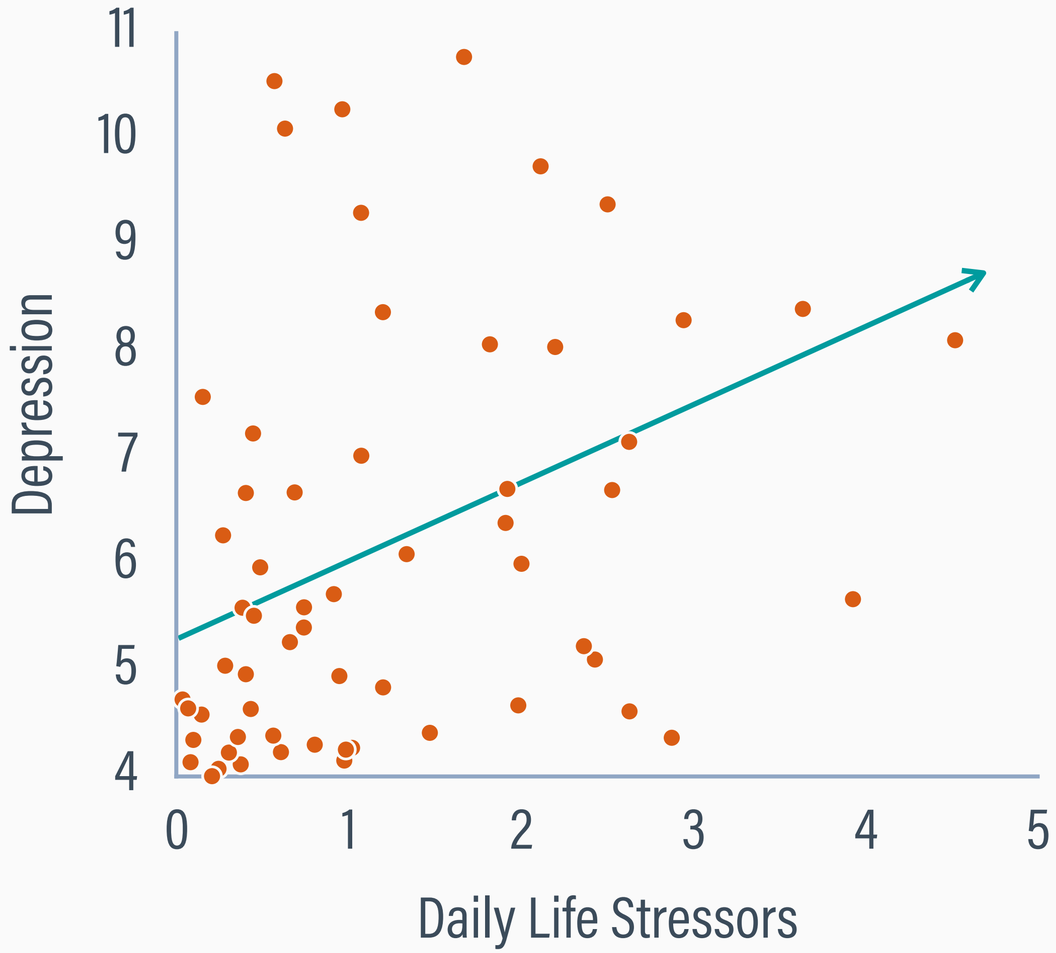


SIGNIFICANCE TESTING STEPS

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

SCATTERPLOT

- There is a positive association: depression increases as daily stressors increase
- A straight line captures the overall positive trend, but the data do not fall on the line
- Pearson's r measures the degree to which the data points fall on a straight line



PRODUCT OF DEVIATION SCORES

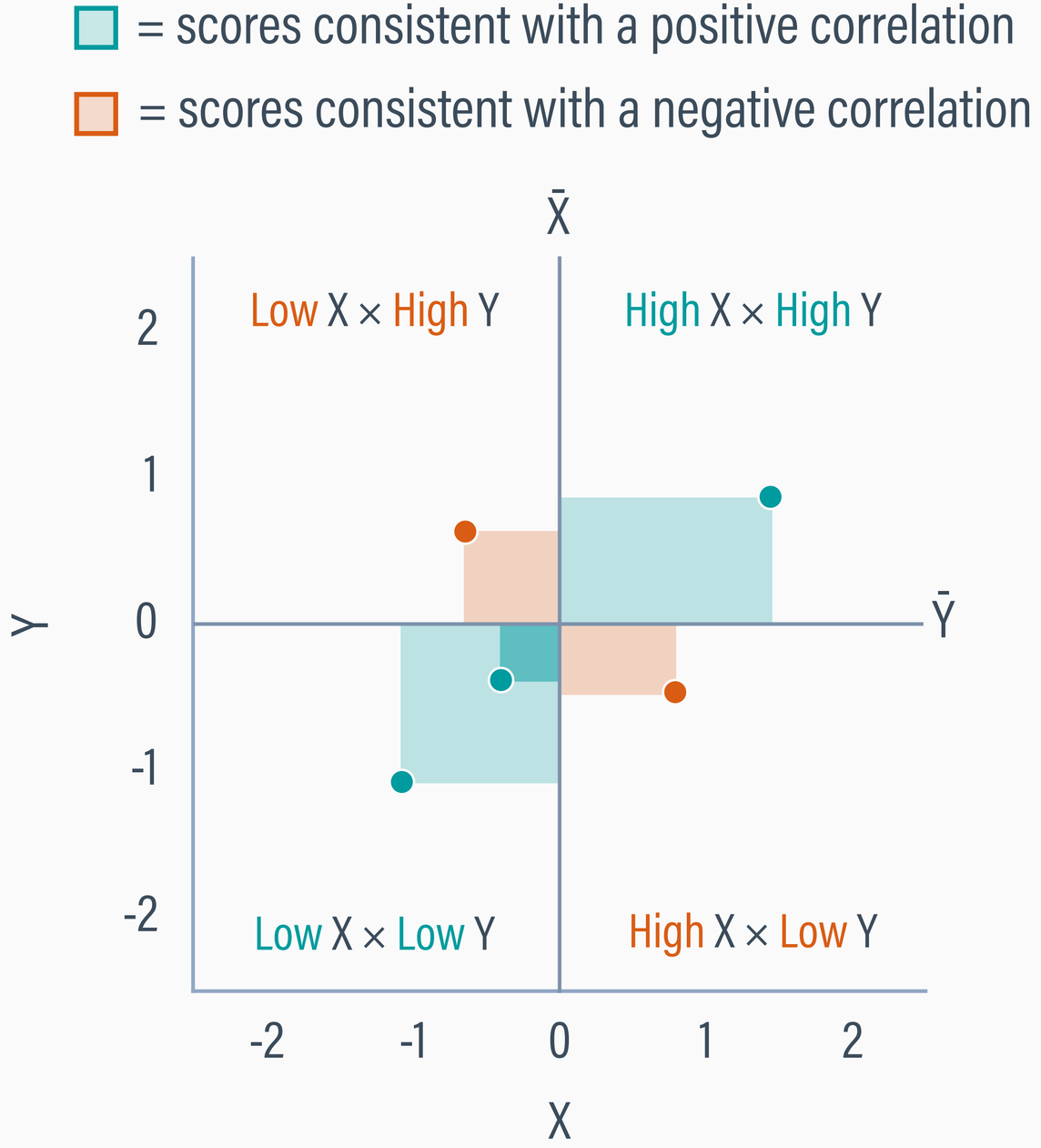
- The product of two deviation scores is the foundation of correlation, as it captures how two variables vary together

$$(X - \bar{X})(Y - \bar{Y}) = (d_X)(d_Y)$$

- A positive product means X and Y increase (or decrease) together, whereas a negative product means that the variables change in opposite directions

DEVIATION PRODUCT AND A SCATTERPLOT

- The product of deviation scores—visualized as the size of the colored squares—encodes the location of scores in the scatterplot
- Positive product values (shown in green) are consistent with a positive trend line
- Negative product values (shown in orange) are consistent with a negative trend line



DEMONSTRATION DATA

X	X̄	d_X	Y	Ȳ	d_Y	$d_X d_Y$
2.8	1.8	1.0	10.0	6.7	3.3	3.17
1.0	1.8	-0.8	4.2	6.7	-2.5	2.10
0.1	1.8	-1.7	4.1	6.7	-2.6	4.55
3.9	1.8	2.1	7.0	6.7	0.3	0.56
0.7	1.8	-1.1	5.6	6.7	-1.1	1.28
2.6	1.8	0.8	4.6	6.7	-2.1	-1.64
0.5	1.8	-1.3	7.5	6.7	0.8	-1.02
3.2	1.8	1.4	6.7	6.7	-0.0	-0.04
1.8	1.8	-0.0	8.1	6.7	1.4	-0.04
1.7	1.8	-0.1	9.5	6.7	2.8	-0.36

SUM OF PRODUCTS

- The sum of products summarizes the the degree to which two variables covary as a lump sum (like sum of squares)

$$SP = \sum(X - \bar{X})(Y - \bar{Y}) = \sum(d_X)(d_Y)$$

- Whether correlation is positive or negative depends on whether the positive products outweigh the negative products or vice versa

DEMONSTRATION DATA

The association is positive because the positive products outweigh the negative products (SP > 0)

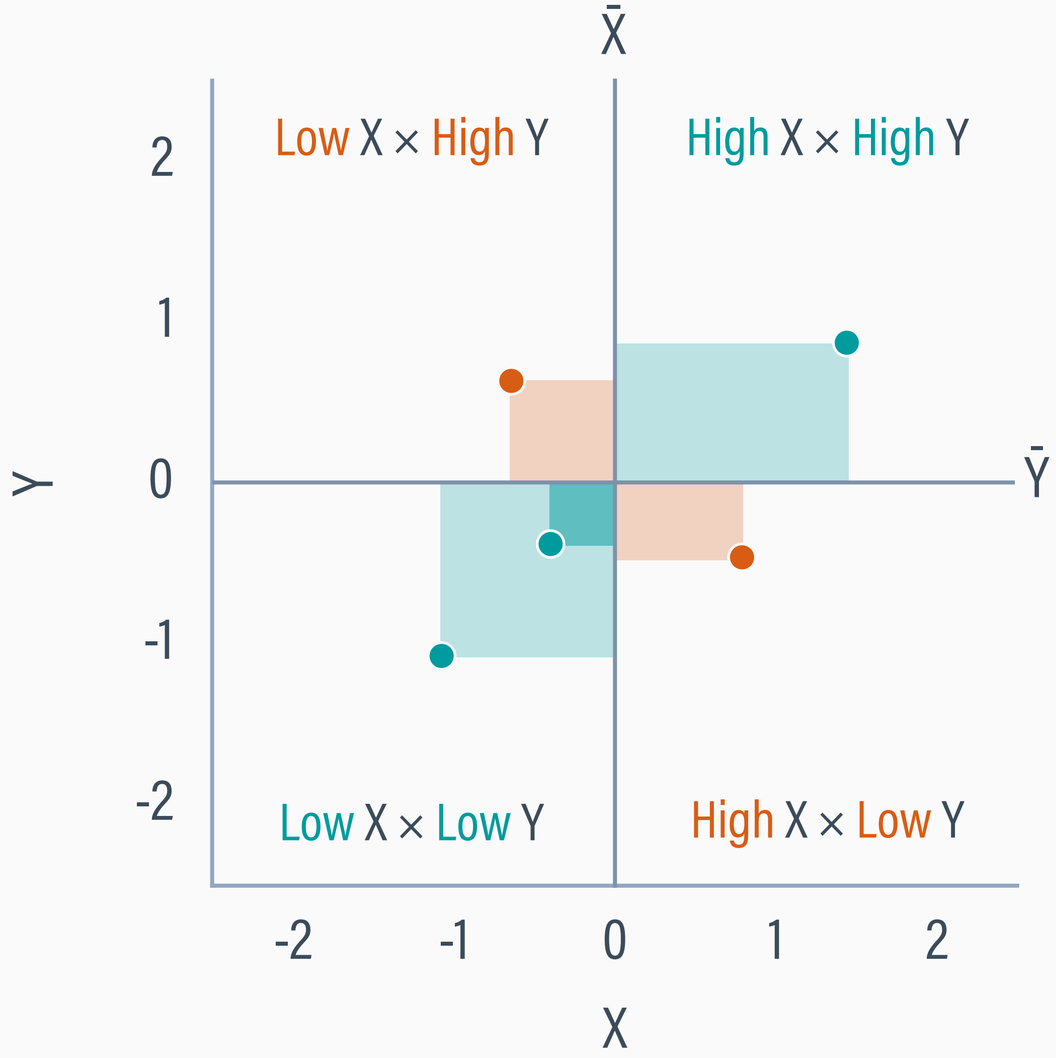
X	X̄	d_X	Y	Ȳ	d_Y	$d_X d_Y$
2.8	1.8	1.0	10.0	6.7	3.3	3.17
1.0	1.8	-0.8	4.2	6.7	-2.5	2.10
0.1	1.8	-1.7	4.1	6.7	-2.6	4.55
3.9	1.8	2.1	7.0	6.7	0.3	0.56
0.7	1.8	-1.1	5.6	6.7	-1.1	1.28
2.6	1.8	0.8	4.6	6.7	-2.1	-1.64
0.5	1.8	-1.3	7.5	6.7	0.8	-1.02
3.2	1.8	1.4	6.7	6.7	-0.0	-0.04
1.8	1.8	-0.0	8.1	6.7	1.4	-0.04
1.7	1.8	-0.1	9.5	6.7	2.8	-0.36

Sum of products = **8.55**

COVARIANCE: UNSTANDARDIZED CORRELATION

- The sum of products expresses the association as a lump sum
- The covariance is the average deviation score product (visually, the average size of the green and orange shaded boxes)

$$s_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N - 1} = \frac{\text{sum of products}}{\text{adjusted sample size (df)}}$$



DEMONSTRATION DATA

- The covariance is positive (as stressors increase, so too does depression)

$$s_{XY} = \frac{SP}{N-1} = \frac{8.55}{10-1} = 0.95$$

- It's strength is hard to assess because the covariance depends on the scales of X and Y (covariance = "raw" correlation)

PEARSON CORRELATION

- Pearson's correlation is the covariance between two z-score variables (a standardized association)

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2} = \frac{SP}{(SS_X)(SS_Y)} = \frac{s_{XY}}{(s_X)(s_Y)}$$

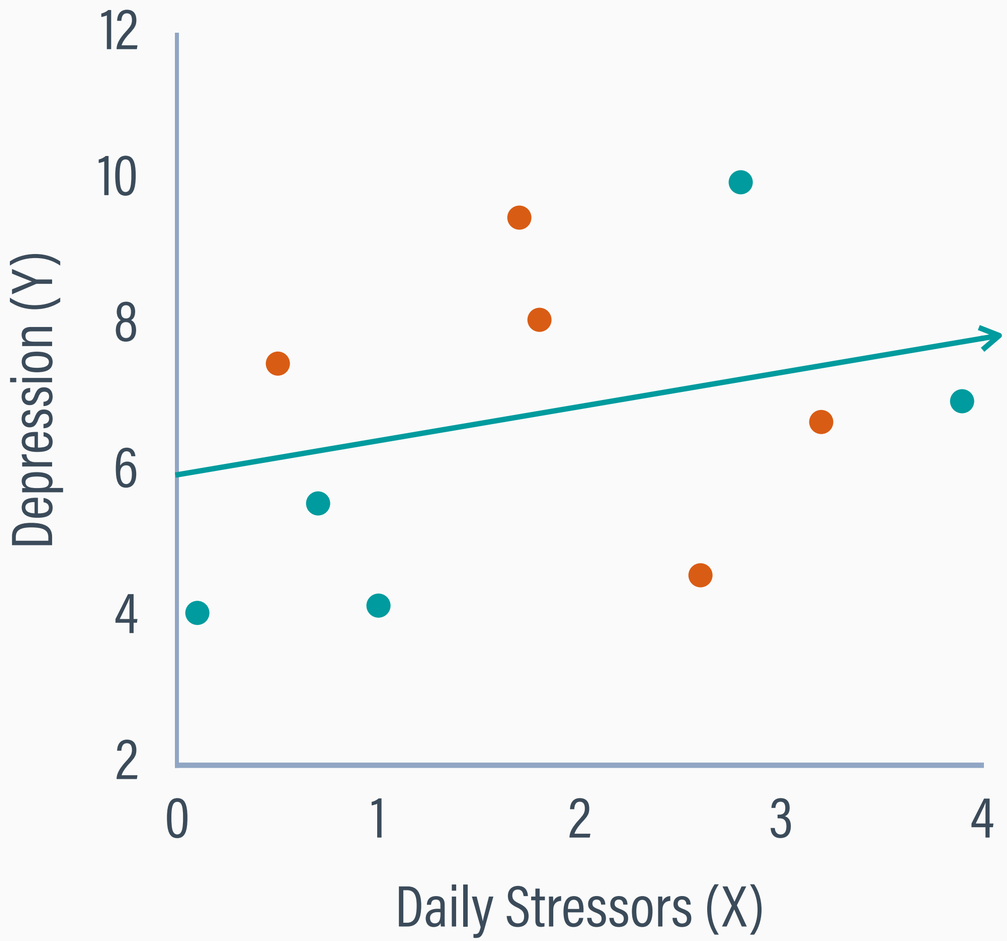
- Standardizing to z-scores removes the influence of scale (e.g., stressors points vs. depression points), forcing r to stay within –1 and +1

DEMONSTRATION DATA

- Dividing the covariance (s_{XY}) by the standard deviations (s_X and s_Y) standardizes the association

$$r = \frac{s_{XY}}{(s_X)(s_Y)} = \frac{0.95}{(1.27)(2.11)} = .35$$

- r = .35 is a positive association and moderate by Cohen's (1988) effect size conventions



SIGNIFICANCE TESTING STEPS

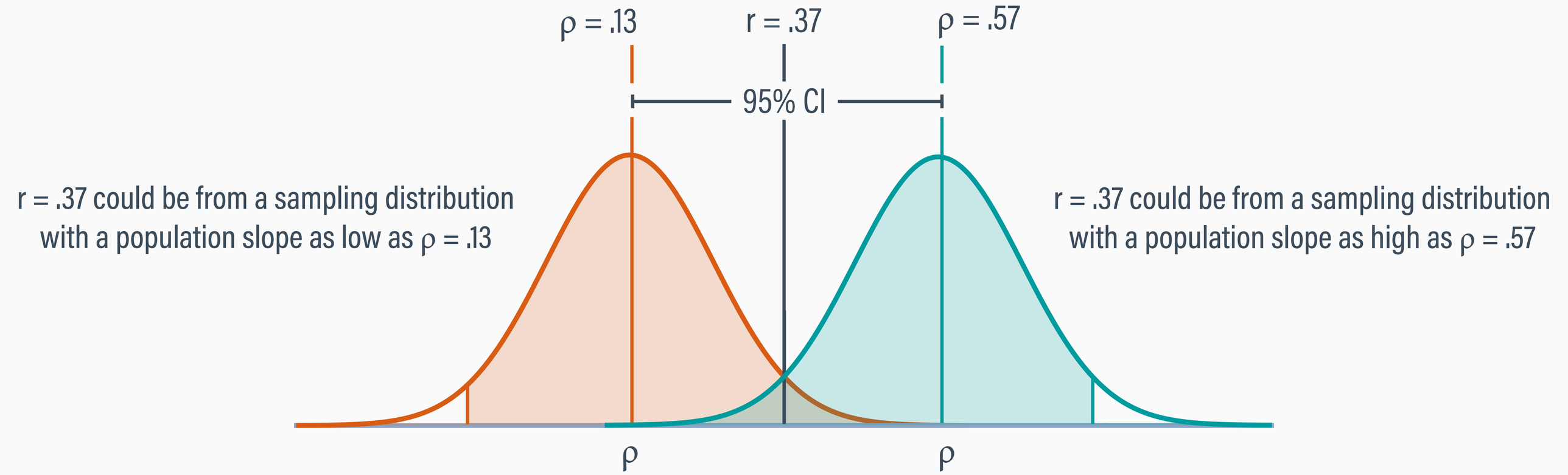
1. Specify hypotheses
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

COMPARING DATA TO THE NULL

- Correlations have skewed sampling distributions, so significance testing transforms the correlation to a z-score, then uses a normal curve for p-values and 95% intervals
- Software programs tend not to report test statistics
- Some report 95% confidence intervals, all report p-values

95% CONFIDENCE INTERVAL

- The 95% confidence interval gives the two most extreme values of the population correlation that could have reasonably produced the sample r



TWO-TAILED P-VALUE

- The p-value tells how likely it is to draw a sample correlation at least as extreme as ours from a null population with $\rho = 0$
- The probability of drawing a sample from the null population with a correlation of at least $\pm.37$ is $p = .003$ (3 out of 1000)
- Only 3 out of every 1000 hypothetical samples from a null population would have correlations this large

SIGNIFICANCE TESTING STEPS

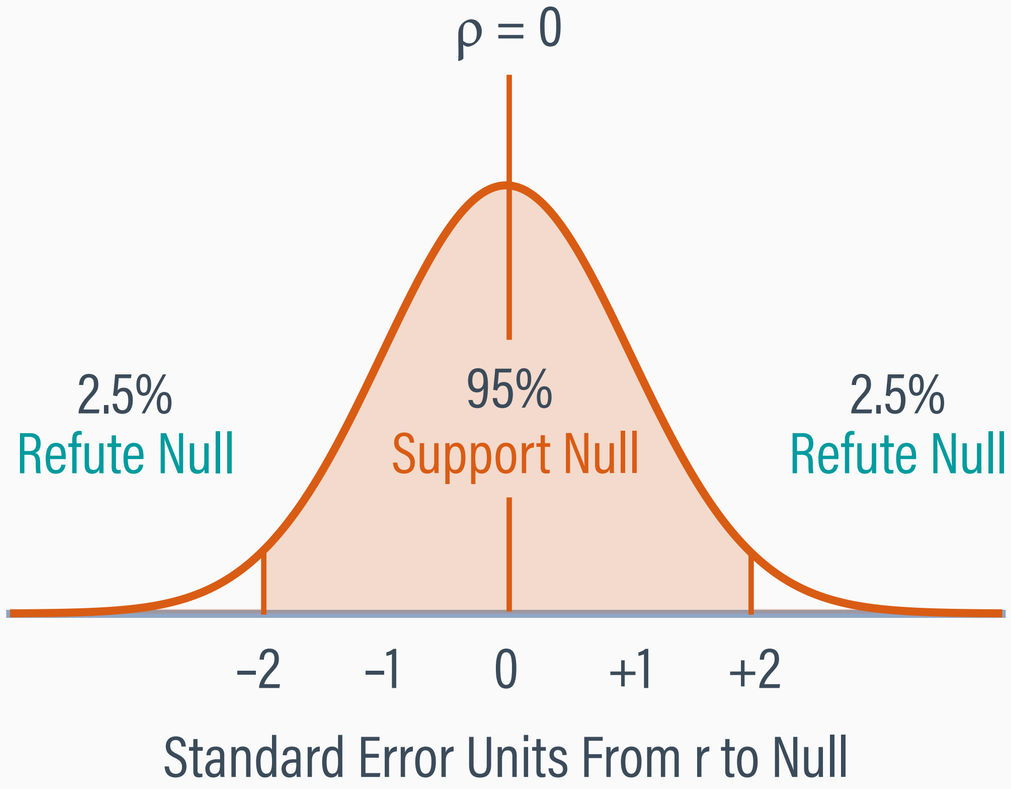
1. Specify hypotheses
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

RESEARCH QUESTION REVISITED

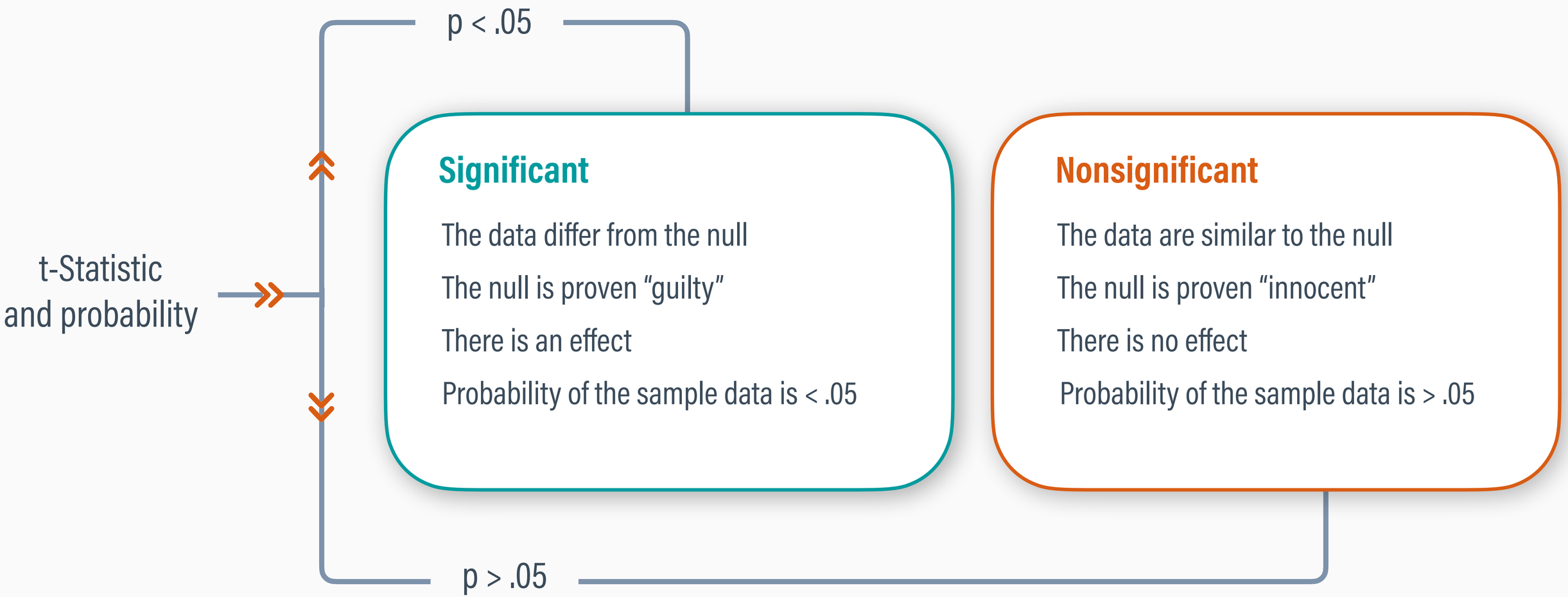
- Studies typically attempt to answer a handful of research questions involving associations between key variables
- Is there an association between daily life stressors and daily depressive symptoms?
- The null hypothesis states that there is no association between life stressors and depression (a flat slope)

5% SIGNIFICANCE CRITERION REVISITED

- By convention, we refute the null if the sample correlation r falls outside the middle 95% of the sampling distribution
- Such as sample has less than a 5% chance of originating from the null population
- We deem the null implausible because our data are unlikely to originate from that population



DECISION TREE



CONCLUSION: TWO-TAILED ALTERNATE

- The p-value of .003 (3 out of 1000) would lead us to refute the null
- A sample correlation as large as $r = \pm.37$ is very unlikely to have originated from a null population with $\rho = 0$
- There is evidence that an increase in daily stressors is associated with a concurrent increase in depression

The two-tailed probability for the study is p = .003. Suppose the researchers had instead specified a one-tailed test where the predicted a positive association (i.e., increases in stressors could only lead to increases in depression). In small groups of two or three, discuss how the p-value would change with a one-tailed alternate hypothesis. Would your conclusion about significance change or stay the same?

OUTLINE

1. Correlation overview
2. Significance testing steps
3. **Correlation matrices for multiple variables**
4. Statistical assumptions
5. Study questions

CORRELATION MATRIX

- Researchers often examine correlations among multiple variables at the same time
- Correlations are arranged in a symmetric matrix with each variable listed in a row and a column
- Each correlation appears twice
- 1s on the diagonal reflect the correlation between a variable and itself

	X	Y	Z
X	1.0	.35	.11
Y	.35	1.0	.21
Z	.11	.21	1.0

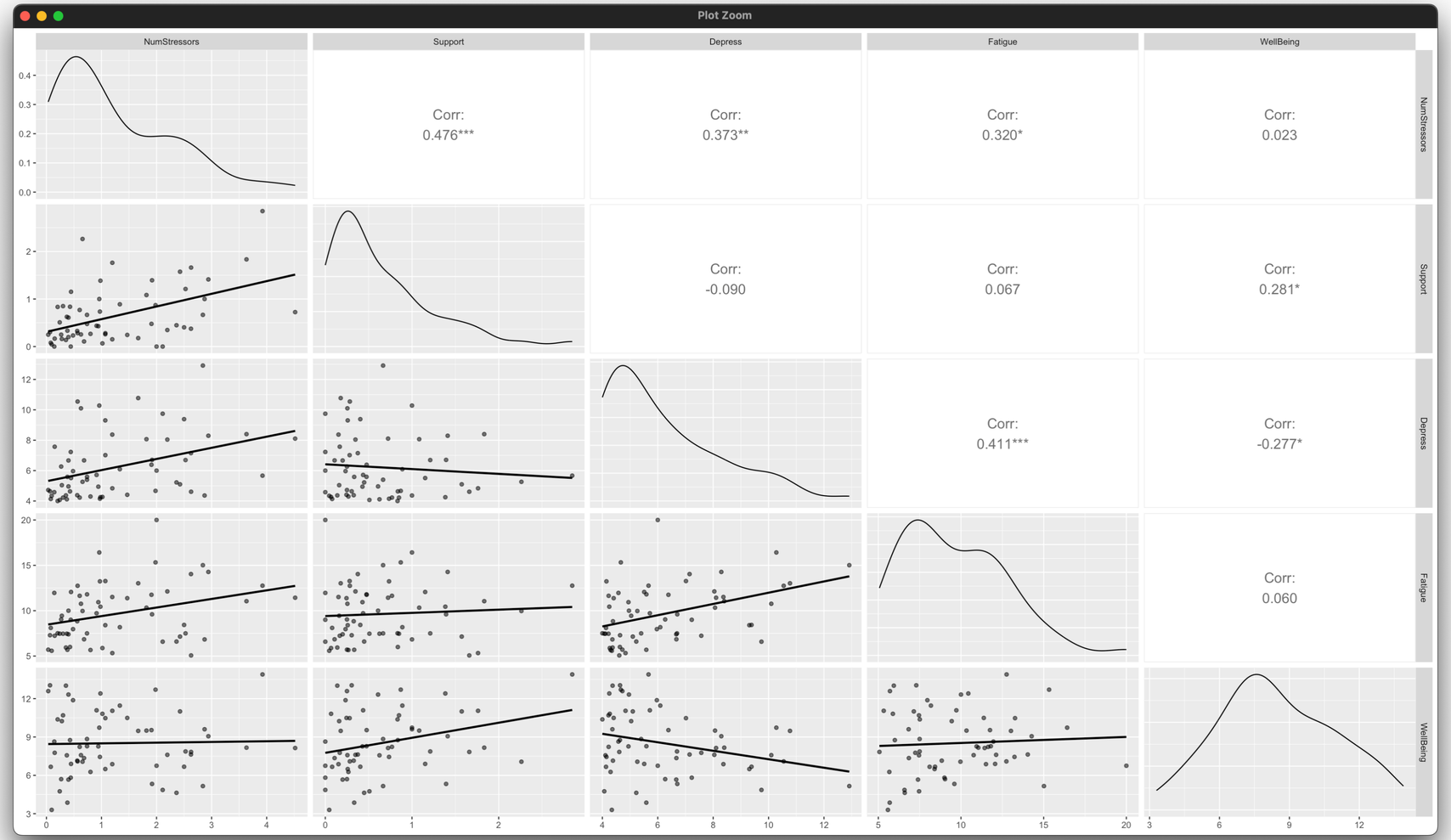
DAILY DIARY DATA

- Researchers collected several daily measures: stressful life events, social support, depression, fatigue, and well-being

	Stressors	Support	Depression	Fatigue	Well-being
Stressors	1.00	.48	.37	.32	.02
Support	.48	1.00	–.09	.07	.28
Depression	.37	–.09	1.00	.41	–.28
Fatigue	.32	.07	.41	1.00	.06
Well-being	.02	.28	–.28	.06	1.00

In small groups of two or three, provide an interpretation of the correlation between social support and depression. Characterize its direction and strength, and describe the trend in a way that a lay person could understand. Do the same for fatigue and depression.

R OUTPUT


Plot Zoom
NumStressors
Support
Depress
Fatigue
WellBeing
Corr:
0.476***
Corr:
0.373**
Corr:
0.320*
Corr:
0.023
Corr:
-0.090
Corr:
0.067
Corr:
0.281*
Corr:
0.411***
Corr:
-0.277*
Corr:
0.060

R OUTPUT

```
Correlation matrix
             NumStressors Support Depress Fatigue WellBeing
NumStressors         1.00    0.48    0.37    0.32      0.02
Support              0.48    1.00   -0.09    0.07      0.28
Depress              0.37   -0.09    1.00    0.41     -0.28
Fatigue              0.32    0.07    0.41    1.00      0.06
WellBeing            0.02    0.28   -0.28    0.06      1.00

Sample Size
[1] 61

Probability values (Entries above the diagonal are adjusted for multiple tests.)
             NumStressors Support Depress Fatigue WellBeing
NumStressors         0.00    0.00    0.02    0.08      1.00
Support              0.00    0.00    1.00    1.00      0.17
Depress              0.00    0.49    0.00    0.01      0.17
Fatigue              0.01    0.61    0.00    0.00      1.00
WellBeing            0.86    0.03    0.03    0.65      0.00
```

R OUTPUT

```
 Confidence intervals based upon normal theory.  To get bootstrapped values, try
cor.ci
            raw.lower raw.r raw.upper raw.p lower.adj upper.adj
NmStr-Spprt      0.25  0.48       0.65  0.00      0.15      0.71
NmStr-Dprss      0.13  0.37       0.57  0.00      0.03      0.64
NmStr-Fatig      0.07  0.32       0.53  0.01     -0.02      0.59
NmStr-WllBn     -0.23  0.02       0.27  0.86     -0.23      0.27
Spprt-Dprss     -0.33 -0.09       0.17  0.49     -0.40      0.23
Spprt-Fatig     -0.19  0.07       0.31  0.61     -0.24      0.36
Spprt-WllBn      0.03  0.28       0.50  0.03     -0.06      0.56
Dprss-Fatig      0.18  0.41       0.60  0.00      0.07      0.66
Dprss-WllBn     -0.49 -0.28      -0.03  0.03     -0.55      0.05
Fatig-WllBn     -0.20  0.06       0.31  0.65     -0.23      0.34
```

OUTLINE

1. Correlation overview
2. Significance testing steps
3. Correlation matrices for multiple variables
4. Statistical assumptions
5. Study questions

STATISTICAL ASSUMPTIONS

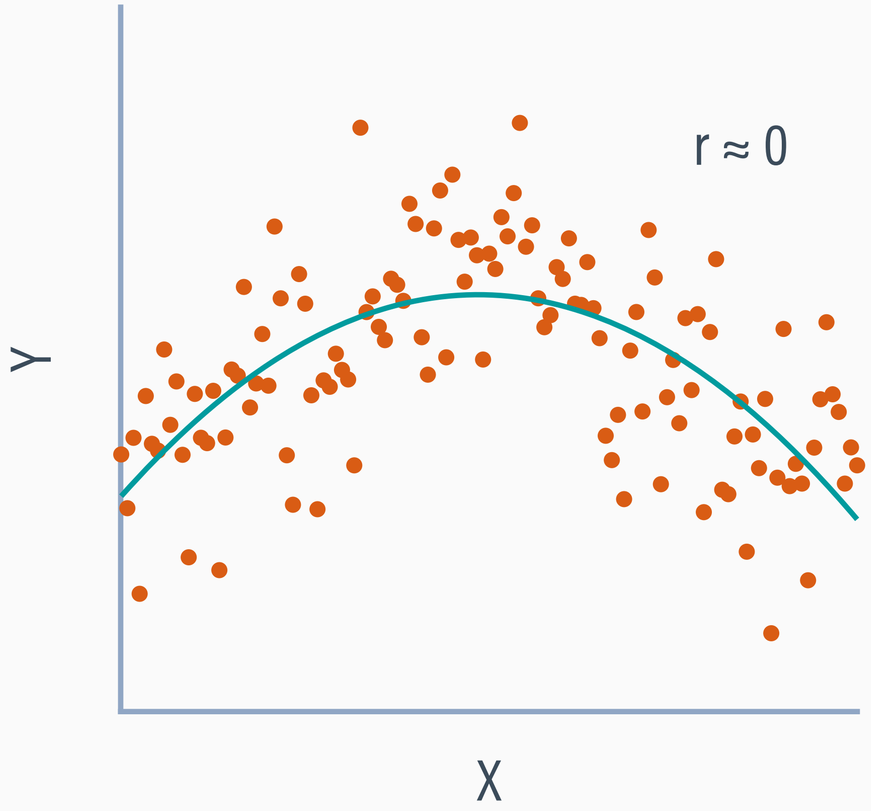
- The accuracy of t-tests (and other statistics) depends on certain conditions in the data being true (e.g., normality)
- Violations of assumptions can bias estimates, inflate or deflate standard errors, and distort significance tests
- Always check reasonableness of assumptions before drawing conclusions

CORRELATION ASSUMPTIONS

- Numeric (approximately continuous) dependent variable
- Association should be linear
- Variables should be normal
- Independence of observations (no participant's score influences any other participant's score)

LINEARITY

- Pearson's correlation is capturing the degree to which data points following a straight line
- Many associations in psychology exhibit nonlinear trends
- Nonlinear trends tend to attenuate the correlation toward zero



NORMALITY

- Both variables should be normally distributed for the z-conversion used in significance testing to be trustworthy
- In small samples, normality violations can artificially inflate or deflate standard errors, thus distorting significance tests
- Normality is less of a concern if the sample size is large enough

OUTLINE

1. Correlation overview
2. Significance testing steps
3. Correlation matrices for multiple variables
4. Statistical assumptions
5. Study questions

STUDY QUESTIONS (1)

1) Generate or find an example of a positive correlation from a psychology study.

2) Generate or find an example of a negative correlation from a psychology study.

3) Within a given study, the researchers use correlation to answer one research question, and they use an independent t-test for another. What features of the variables would cause them to adopt different analyses for the two questions?

STUDY QUESTIONS (2)

4) Researchers use a correlation to determine whether there is an association between one's level of visual impairment and their optimism. The correlation is -.35. Provide an interpretation of the correlation between social support and depression. Characterize its direction and strength, and describe the trend in a way that a lay person could understand.

5) The probability value for the correlation is .12 (12%). Provide an interpretation of the p-value. Discuss your decision about the the null hypothesis. Translate your decision into a tangible statement about the association between daily visual impairment and optimism.