PS 211: Introduction to Experimental Design

Fall 2025 · Section C1

Lecture 16: Correlation

Updates & Reminders

- The Data Write-up was due yesterday.
  - We are going to grade them and post feedback by the end of the week.
  - We will accept late submissions through Friday.
- Homework 4 has been posted.
  - Due **Tuesday, Dec. 2** at 11:59 PM.
  - We are going to post answers on Wednesday, Dec. 3.

Correlation

Recap: Statistical Tests We've Covered So Far

Test	# of IVs	IV Type	# of Levels	DV Type	Use Case
z-test	0	—	—	Numeric	Compare sample mean to population mean (known SD)
One-Sample t-test	0	—	—	Numeric	Compare sample mean to population mean (unknown SD)
Independent t-test	1	Categorical	2	Numeric	Compare two groups (e.g., School A vs. School B)
Paired t-test	1	Categorical	2	Numeric	Compare same group before vs. after intervention or in two conditions
One-Way (Between-Groups) ANOVA	1	Categorical	3+	Numeric	Compare 3+ groups (e.g., Drug A, B, C)
Repeated-Measures (Within-Subjects) ANOVA	1	Categorical	3+	Numeric	Compare same participants across 3+ conditions

Recap: Statistical Tests

Test	# of IVs	IV Type	# of Levels	DV Type	Use Case
z-test	0	—	—	Numeric	Compare sample mean to population mean (known SD)
One-Sample t-test	0	—	—	Numeric	Compare sample mean to population mean (unknown SD)
Independent t-test	1	Categorical	2	Numeric	Compare two groups (e.g., School A vs. School B)
Paired t-test	1	Categorical	2	Numeric	Compare same group before vs. after intervention or in two conditions
One-Way (Between-Groups) ANOVA	1	Categorical	3+	Numeric	Compare 3+ groups (e.g., Drug A, B, C)
Repeated-Measures (Within-Subjects) ANOVA	1	Categorical	3+	Numeric	Compare same participants across 3+ conditions
Correlation	**1**	**Numeric**	—	**Numeric**	**Examine the relation between two numeric variables**

Defining Correlation

- A **correlation** is a systematic association (relation) between two variables.
- It reflects **covariation** ("co-relation") — how two numeric variables change together.
- As discussed earlier in the course, correlations do *not* tell us about causation!
- We typically use correlations in **observational** (non-experimental) research designs.

Review: Correlation vs. Causation

- **Correlation:** An association between two or more variables
  - Variables are typically measured, not manipulated.
  - Shows relations — but **not cause.**

Advantages

- Some research questions **cannot be studied experimentally** (e.g., unethical to manipulate the amount of stress in children's early life environments).
- Efficient way to study **naturally occurring variables.**

Disadvantages

- **Confounding variables** can create false associations.
- **Causality cannot be inferred** — we can't tell which variable influences which.

Characteristics of Correlations

- Correlations are summarized by the **Pearson correlation coefficient** (r).
- r indicates both:
  - **Direction** (positive or negative) of the relationship.
    - Determined by the sign of r (+ or –).
  - **Strength** (magnitude) of the relationship.
    - Determined by the absolute value of r (ignoring sign).

Interpreting r values:

- r ranges from –1 to +1.
- $r = 0$ → no linear relation

Assumptions of Pearson's *r*

- Variables are **numeric**.
- The relation between two variables is **linear**. (Pearson's *r* only captures linear relationships.)
- Scores have been **randomly sampled** from the population.
- The distributions from which the scores have been sampled are **approximately normal**.

Directions of Correlation

- **Positive correlation:**

  When one variable increases, the other tends to **increase** too.

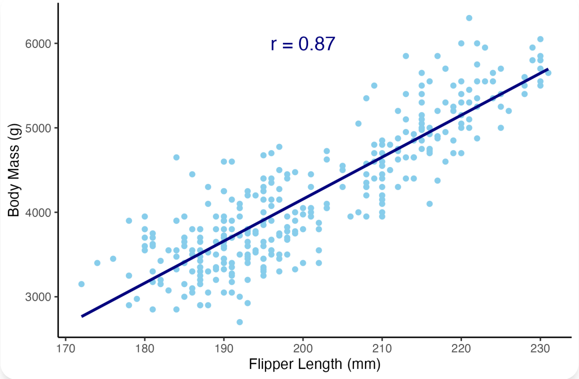
  > Example: Infants who hear more words in the first year tend to have larger vocabularies at age 2 (r = .52).

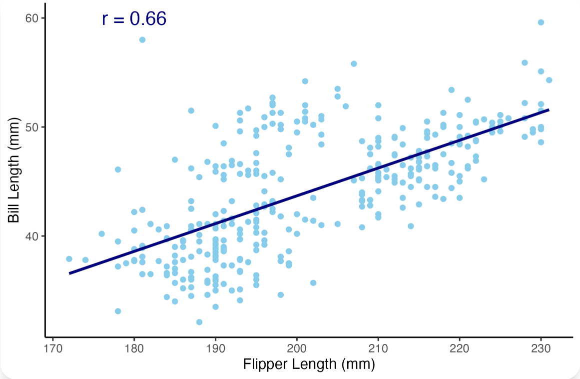
- **Negative correlation:**

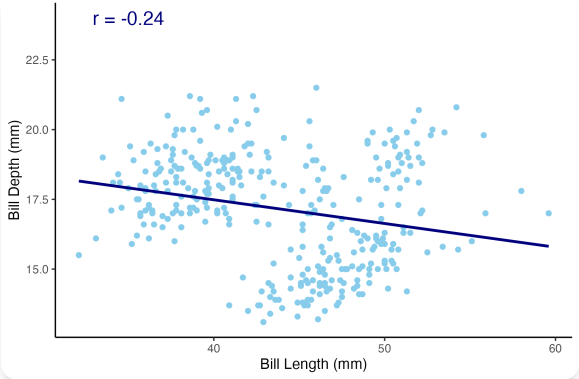
  When one variable increases, the other tends to **decrease**.

  > Example: Students who cheat more on exams tend to have lower grades overall (r = −0.43).

Examples: Penguin dimensions







Practice: Identify the Correlation Type

Which of the following would you expect to show a **negative correlation**?

A. Peoples' heights and weights

B. The number of hours people spend studying and their exam scores

C. Distance people live from campus and their attendance in classes

D. Shoe size and GPA

Answer: C. We would expect a negative correlation between distance from campus and class attendance — as distance increases, attendance tends to decrease.

A. We would expect a *positive* correlation between height and weight — as height increases, weight tends to increase as well.

B. We would expect a *positive* correlation between the number of hours people spend studying and their exam scores — as study time increases, exam scores tend to increase as well.

D. We would expect *no correlation* between shoe size and GPA — these variables are unrelated.

Correlation Strength

- The **magnitude** of *r* (ignoring its sign) shows **strength**.
- Larger absolute *r* → stronger linear relationship.

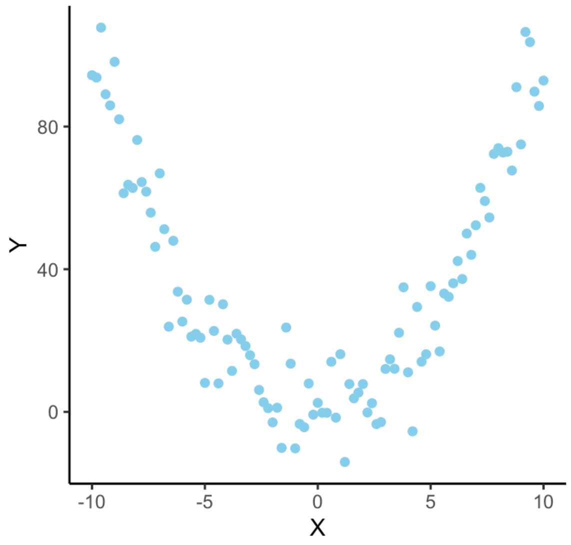
Cohen's (1988) general guidelines:

Strength	|r| value	Interpretation
Small	.10	Weak relationship
Medium	.30	Moderate relationship
Large	.50	Strong relationship

In behavioral science, *r* values above .50 are **rare** — human behavior is noisy!

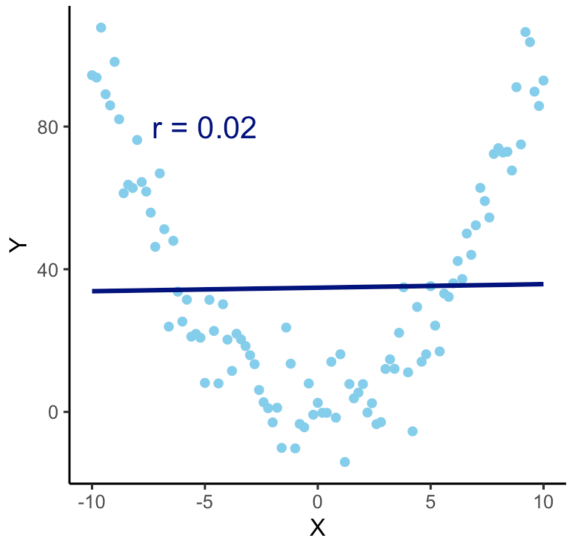
Correlations are *Linear*

What is the correlation between these two variables?



Correlations measure *linear* relationships

What is the correlation between these two variables?



- There is *no* linear relationship here — the pattern is curved.

BOSTON UNIVERSITY

Part 2: Calculating and Interpreting *r*

The Pearson Correlation Coefficient (*r*)

- The Pearson correlation coefficient is the most common measure of correlations between two variables.
- It quantifies the **linear relationship** between two numeric variables.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- x_i, y_i = individual scores
- $\bar{x}$, $\bar{y}$ = means of X and Y

The Pearson Correlation Coefficient (*r*)

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

- Numerator = product of deviations from means (how X and Y move together)
- Denominator = product of sums of squared deviations (standardizes the measure)

The Pearson Correlation Coefficient (*r*): Numerator

Understanding the numerator:

$$\sum(x_i - \bar{x})(y_i - \bar{y})$$

- For each pair of scores, calculate how far each score is from its mean.
- Multiply these deviations together.
- Sum these products across all pairs.
- Positive products (both above or both below means) increase *r*.
- Negative products (one above, one below) decrease *r*.

The Pearson Correlation Coefficient (*r*): Demoninator

Understanding the demoninator:

$$\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}$$

- For each variable, calculate squared deviations from the mean.
- Sum these squared deviations.
- Multiply the sums for X and Y, then take the square root.
- This gives us an overall measure of variability for both variables.
- This standardizes the correlation, ensuring *r* ranges from –1 to +1.
- Like our other statistics, *r* can be understood as a **signal-to-noise ratio**:
  - Numerator = signal (covariation)

From Scatterplots to r

- We can visualize relationships using a **scatterplot**.
  - Each point represents one participant.
  - The participant's score on variable X is on the x-axis; score on variable Y is on the y-axis.
- Scatterplots show:
  - Direction (positive / negative)
  - Strength (tightness of clustering)

Remember: correlation measures **linear** relationships only — curved patterns can have $r \approx 0$ even if related.

The Pearson Correlation Coefficient (r): Understanding visually

The Pearson Correlation Coefficient (r): Understanding visually

Statistical Significance of *r*

- A significant *r* means the relationship between the two variables is **unlikely to be due to chance**.
  - This means that the observed correlation in our sample likely reflects a true relation that exists in the population.
- We can test the significance of *r* using the *t* statistic!
  - This is a little confusing because we have previously used the *t* statistic to compare means, but here we use it to test correlations.
  - This is a totally different test - the only similarity is that we are using the *t* distribution as our null distribution.

For a correlation with sample size *N*, we compute:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Statistical Significance of *r*

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Don't worry about memorizing this formula or understanding where it comes from.

Key ideas:

- Larger absolute r → larger t → more likely to be significant.
- Larger N → larger t → more likely to be significant.
- Larger N → more degrees of freedom → smaller critical t → more likely to be significant.

Coefficient of Determination (R^2)

- R^2 = proportion of shared variance between two variables.
- We discussed R^2 in the context of ANOVA - it can also be used with correlations, and in fact, is directly related to r.

$$R^2 = r^2$$

- R^2 helps us understand how much of the variability in one variable is explained by the other.
- If the two variables are correlated, that means that we can account for some of the variance in one variable by the other variable.
- R^2 ranges from 0 to 1.
- Larger R^2 → more shared variance → stronger relationship.

Example:
If $r = 0.90$, then $R^2 = 0.81$ → 81% of variance in one variable is explained by the other.

The remaining 19 % is due to chance, measurement error, or other factors.

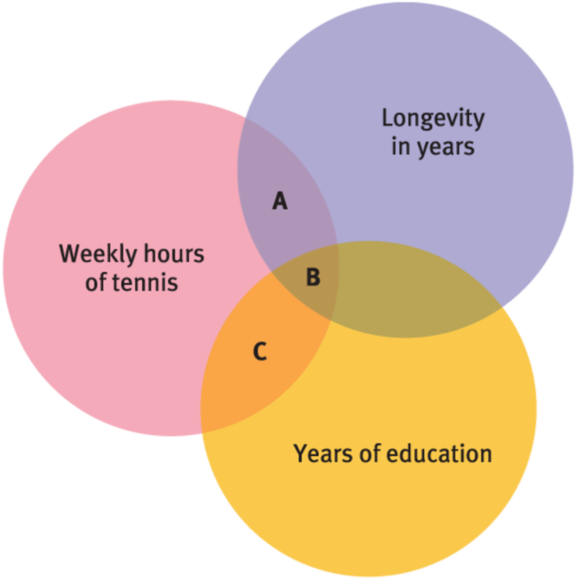
Partial Correlations

- A **partial correlation** shows how much of the relationship between two variables remains after removing the influence of a **third variable**.
- In other words, the correlation coefficient expresses the relationship between two variables, over and above their association with a third variable.

Partial Correlations

Example:

- People who play more sports tend to live longer.
- Some of that relationship may be due to things like smoking, income, or alcohol use.
- When we statistically remove (control for) those variables, the correlation gets smaller.
- But if we remove the effect of **education**, the correlation between sports and longevity is still significant.
- This tells us that **sports and longevity are related above and beyond education** — that remaining relationship is a **partial correlation**.



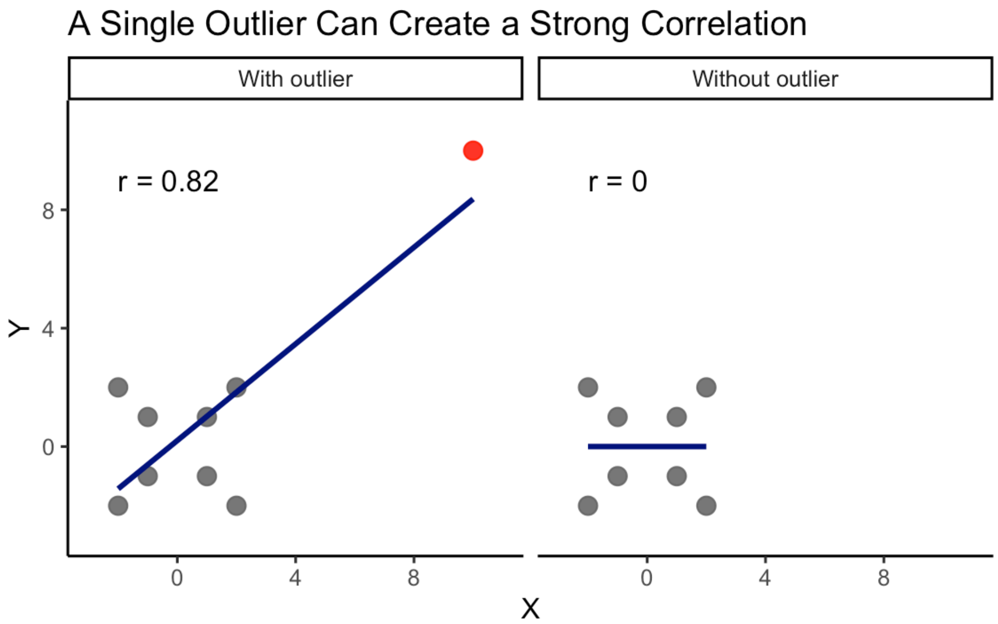
Nolan/Heinzen, *Statistics for the Behavioral Sciences*, 5e, © 2020 Worth Publishers

Outliers and Correlations

- Outliers can **inflate** or **deflate** *r* dramatically.

Outliers and Correlations

- Outliers can **inflate** or **deflate** *r* dramatically.



Defining Outliers

- An **outlier** is a data point that differs significantly from other observations.
- Outliers can arise from:
  - Measurement errors (e.g., data entry mistakes)
  - Natural variability (genuine extreme values)
  - Sampling issues (e.g., non-representative samples)
- There are multiple methods to identify outliers, including visual inspection and statistical tests.
  - It is always extremely important to visually inspect your data before conducting analyses.

Detecting Outliers with Statistical Tests

- Common methods include:
  - **Grubb’s Test:** Identifies a single outlier in a normally distributed dataset.
    - How far the suspected outlier is from the mean compared to the standard deviation.
  - **Dixon Q’s Test:** Suitable for small sample sizes to detect a single outlier.
    - Compares the gap between the suspected outlier and its nearest neighbor to the range of the data.
  - **IQR Method:** Values beyond 1.5 × IQR from Q1 or Q3 are considered outliers.

Outliers: What should you do?

Should you drop outliers?

Drop if:

- The data point is clearly wrong (error, typo).

Keep them if:

- They reflect genuine variation or critical outcomes.

Other ways of dealing with outliers

- Some analysis methods are robust to outliers (e.g., Spearman's rank correlation).
- Sometimes you can report results with and without outliers to show their impact.

Practice: Interpreting Correlations

Hill (1990) studied final exam grades in Sociology and found these correlations:

Variable	*r*
Overall GPA	.72
Number of absences	−.51
Hours spent studying	.31

Which variable shows the **strongest** relationship with exam grade?

A. Hours spent studying
B. Number of absences
C. Overall GPA

Answer: C. Overall GPA (r = .72) has the strongest positive relationship with exam performance.

Practice: Direction & Causation

A study finds a correlation of $r = -.45$ between screen time and sleep quality.
Which interpretation is most appropriate?

A. Screen time causes poor sleep.

B. Poor sleep causes increased screen time.

C. A third variable (e.g., stress) affects both screen time and sleep quality.

D. There is a negative association between screen time and sleep quality.

Answer: D. There is a negative association between screen time and sleep quality. We cannot infer causation from correlation alone.

Part 3: Correlations in R

R Practice: Computing Correlations

```r
# Sample data
df <- data.frame(
  study_hours = c(2, 4, 5, 6, 8, 10),
  exam_score  = c(55, 63, 68, 72, 85, 91)
)
```

Compute Pearson correlation

```r
cor(df$study_hours, df$exam_score)
```

Test significance

```r
cor.test(df$study_hours, df$exam_score)
```

Visualize with ggplot2

```r
library(ggplot2)
ggplot(df, aes(x = study_hours, y = exam_score)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal()
```

That's all for today!

Next class: Regression!