

STAT540: Computing in Statistics

Data Visualization and Confounding

Yen-Yi Ho

Department of Statistics, University of South Carolina

Section 1

Data Visualization

Question 1

Two variables have the **same** median, quartiles and range, so their boxplots look identical. One is bell-shaped; the other has **two separate peaks**.

Which plot would reveal the difference?

- A.** A boxplot with a smaller IQR multiplier
- B.** A density plot or histogram
- C.** A QQ-plot against the Normal
- D.** Nothing — identical summaries mean identical distributions

Question 1

Two variables have the **same** median, quartiles and range, so their boxplots look identical. One is bell-shaped; the other has **two separate peaks**.

Which plot would reveal the difference?

- A.** A boxplot with a smaller IQR multiplier
- B.** A density plot or histogram
- C.** A QQ-plot against the Normal
- D.** Nothing — identical summaries mean identical distributions

Answer: B.

A boxplot reports **five numbers**, and two peaks are invisible to all five. **C** would flag that one is not Normal, but not *why*. **D** is the habit worth breaking: summaries are lossy, which is the whole reason we plot.

Question 2

You want the **colour of the points to depend on** the sex column.

Which line does that?

- A. `geom_point(color = "sex")`
- B. `geom_point(aes(color = sex))`
- C. `geom_point(color = sex)`
- D. `scale_color_manual(sex)`

Question 2

You want the **colour of the points to depend on** the `sex` column.

Which line does that?

- A. `geom_point(color = "sex")`
- B. `geom_point(aes(color = sex))`
- C. `geom_point(color = sex)`
- D. `scale_color_manual(sex)`

Answer: B.

Anything data-driven goes **inside** `aes()`. Outside `aes()` you are *setting* a constant, so **A** paints every point the literal colour “sex” (an error), and **C** looks for an object called `sex` in your workspace, not in the data.

Section 2

Confounding

Question 3

In NHANES, respondents who report **no** physical activity have systolic pressure 2.28 mmHg higher.

Which fact establishes that age could confound this comparison?

- A.** Inactive respondents are 4.6 years older on average
- B.** Blood pressure rises with age ($r = 0.36$)
- C.** Both A and B
- D.** Neither — age is on the causal pathway

Question 3

In NHANES, respondents who report **no** physical activity have systolic pressure 2.28 mmHg higher.

Which fact establishes that age could confound this comparison?

- A. Inactive respondents are 4.6 years older on average
- B. Blood pressure rises with age ($r = 0.36$)
- C. Both A and B
- D. Neither — age is on the causal pathway

Answer: C — you need both.

A confounder must be associated with **the grouping** *and*, independently, with **the outcome**. Either one alone is not enough. (**D** describes a *mediator* — and activity does not cause your age.)

Question 4

The crude difference was 2.28 mmHg. Adjusted for age, it is 0.50 mmHg.

Why is the crude number so much larger?

- A. The crude difference is not a weighted average of the decades
- B. Each group's mean is weighted by **its own** age distribution
- C. Some decades have too few people to be reliable
- D. Rounding error accumulated across the six decades

Question 4

The crude difference was 2.28 mmHg. Adjusted for age, it is 0.50 mmHg.

Why is the crude number so much larger?

- A. The crude difference is not a weighted average of the decades
- B. Each group's mean is weighted by **its own** age distribution
- C. Some decades have too few people to be reliable
- D. Rounding error accumulated across the six decades

Answer: B.

The inactive group carries 11.7% of its weight in the 70+ decade; the active group carries 5.3%. Two weighted averages over **different populations**. Adjustment forces one common weight vector onto both.

Q	Idea	Answer
1	Summaries are lossy — a boxplot hides two peaks	B
2	Data-driven $\Rightarrow$ inside <code>aes()</code>	B
3	A confounder needs <i>both</i> associations	C
4	Two groups, two weight vectors	B

Compare like with like — then say which “like” you chose.