Econ 270 Lecture 3

Sam Gifford

2024-01-22

Two Variable Summaries

- Often we want to look at how different variables relate to each other
- Like univariate analysis, there are both broad ways to assess this and simpler summary statistics

Contingency Tables (2.2.1)

- Also called crosstabulation
- Analog of a histogram: given values of two separate variables, how many are in each cell?
- Commonly implemented in Excel using pivot tables

Contingency Tables

- Suppose I'm looking at exam scores and attendance
- One simple thing I can do is split each into 'above median' and 'below median'
- This creates 4 possible categories

Contingency tables - scores

	Low Grade	High Grade	Total
Low Attendance	6	4	10
High Attendance	3	7	10
Total	9	11	20

- Note that cells where attendance and exam scores agree are more populated
- Can look across rows or down columns for marginal analysis

Marginal Analysis

- What percent of students had low attendance?

	Low Grade	High Grade	Total
Low Attendance	6	4	10
High Attendance	3	7	10
Total	9	11	20

Joint Analysis

- What percent of students had low attendance but received a high grade?

	Low Grade	High Grade	Total
Low Attendance	6	4	10
High Attendance	3	7	10
Total	9	11	20

Conditional Analysis

- Among students with low attendance, what percentage received a low exam grade?

	Low Grade	High Grade	Total
Low Attendance	6	4	10
High Attendance	3	7	10
Total	9	11	20

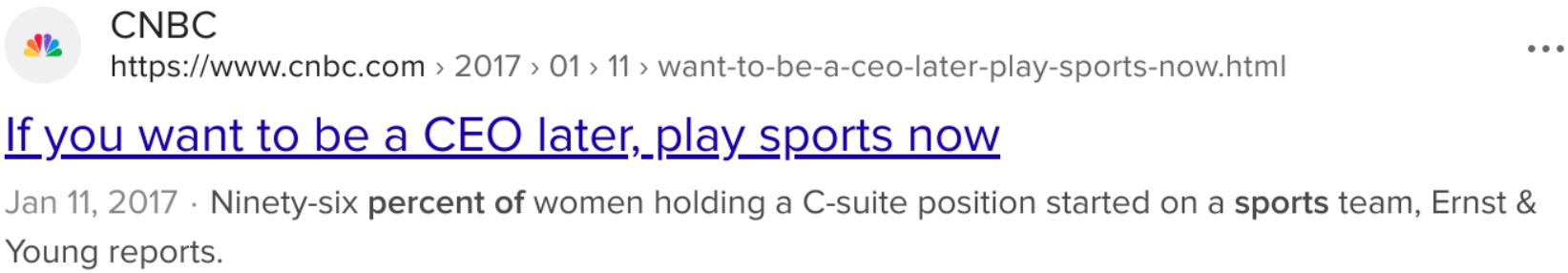
Misinterpreting Statistics

- What percent of people who play sports are executives?
- What percent of executives play sports?

	Play	Don't	Total
Exec	2.4	0.1	2.5
Non-Exec	78.0	19.5	97.5
Total	80.4	19.6	100.0

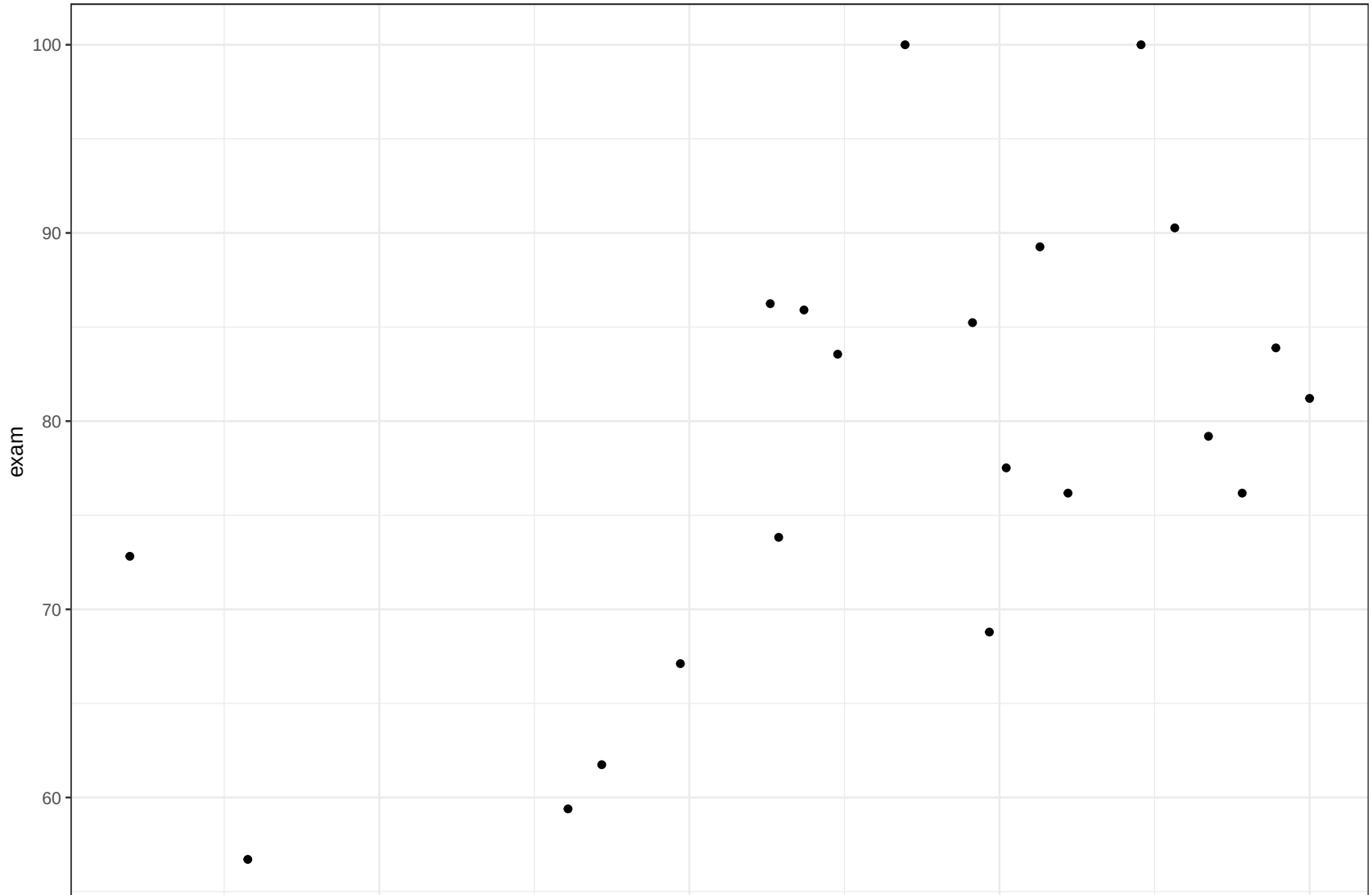
Misinterpreting Statistics

- Data from BLS, the Sport and Fitness Industry Association, and Korn Ferry. Data values rounded.



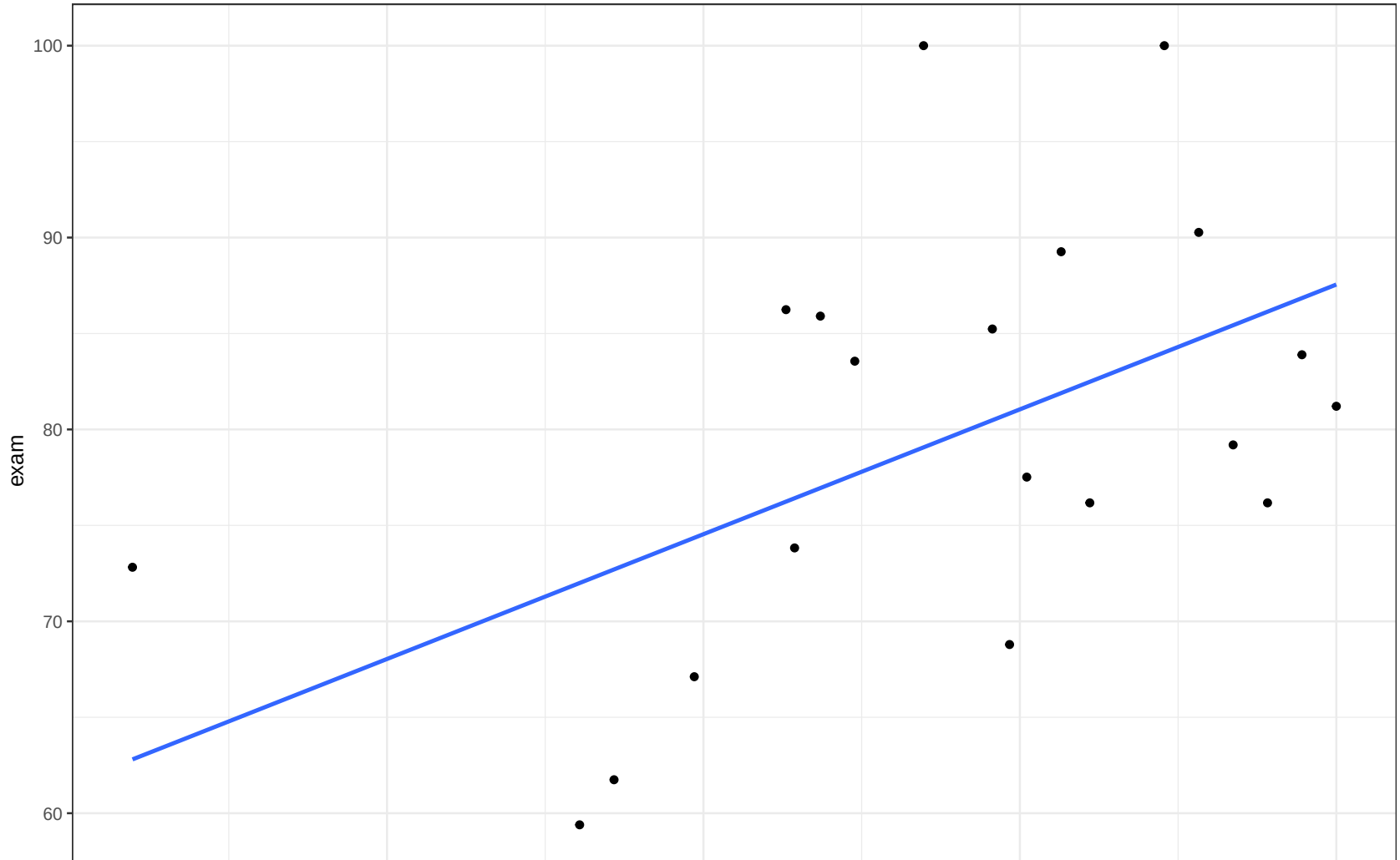
scatterplots (2.1.1)

- If we have quantitative data, we can display two variables on a standard x-y plane
- The corresponding graph is called a scatterplot



Trendlines

- It is common to add a line of best fit, or trendline. You will calculate how to do this in econ300. The best fit line minimizes the sum of squared distances from the line of best fit



positive/negative/uncorrelated

- In the prior graph, we saw that as students with higher attendance also had higher exam scores, on average
- If we can fit a positive line to data there is a positive association, or correlation between variables
- A negative sloping line has a negative association
- If the slope is zero, the variables are uncorrelated

correlation as measure

- Correlation ranges from -1 to 1. The sign tells us whether there is a positive or negative association. The absolute value tells us the strength
- r=1 means a perfect positive association. r=-1 means a perfect negative association. r=0 means no linear association.

Correlation Examples

https://mcfortran.shinyapps.io/correlation/

Correlation Calculation

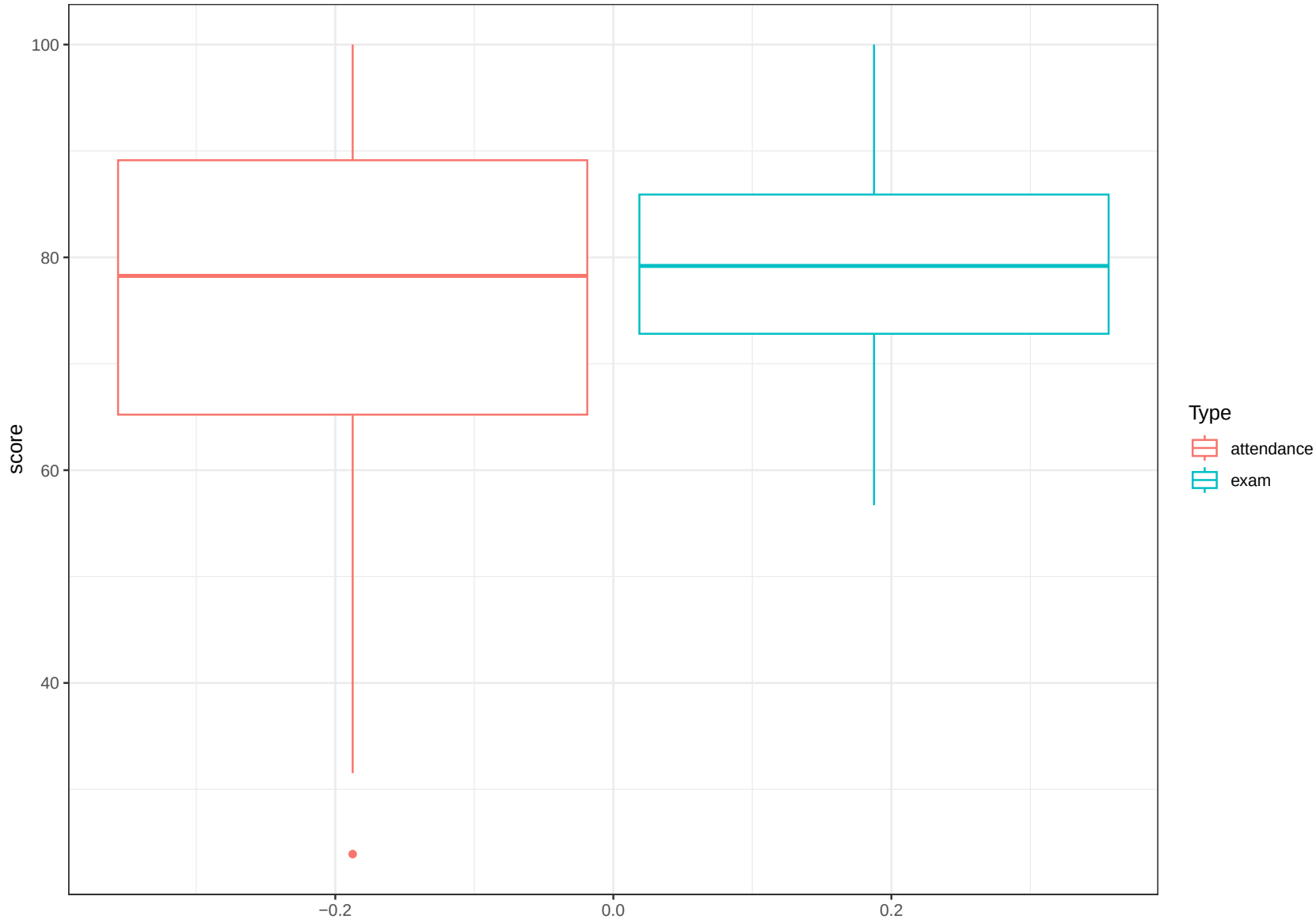
- We likely won't calculate correlation in this class, but it is actually very similar to variance
- $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
- When x and y are both above or below the mean it contributes positively. When one is above and the other is below it contributes negatively.
- Not robust! Outliers in both the x- and y- direction have large influence ('leverage')

A note on regression fit

- The correlation coefficient, r, tells us how close a line fits the data
- It turns out that r^2 actually has a nice interpretation: it's the percent of the variation in y that is explained by x
  - So $r = .7$ means that our regression can explain half of the total variation
- This sometimes gets glossed over in econometrics, but it's something you should know.

Overlayed histograms/boxplots (2.2.6)

- We can also show two histograms or box plots side-by-side to compare distributions



Overlayed histograms/boxplots (2.2.6)