March 9

- Today
  - Chapter 6: Correlation
  - Beginning of Chapter 7: Linear Regression
- Tomorrow
  - Practical: to be posted this afternoon
- For next week
  - Homework: to be posted later today
  - Read Chapter 7
- 1.265708e+00
- replacing outliers with mean +/- 2 standard deviations

Correlation

Aims

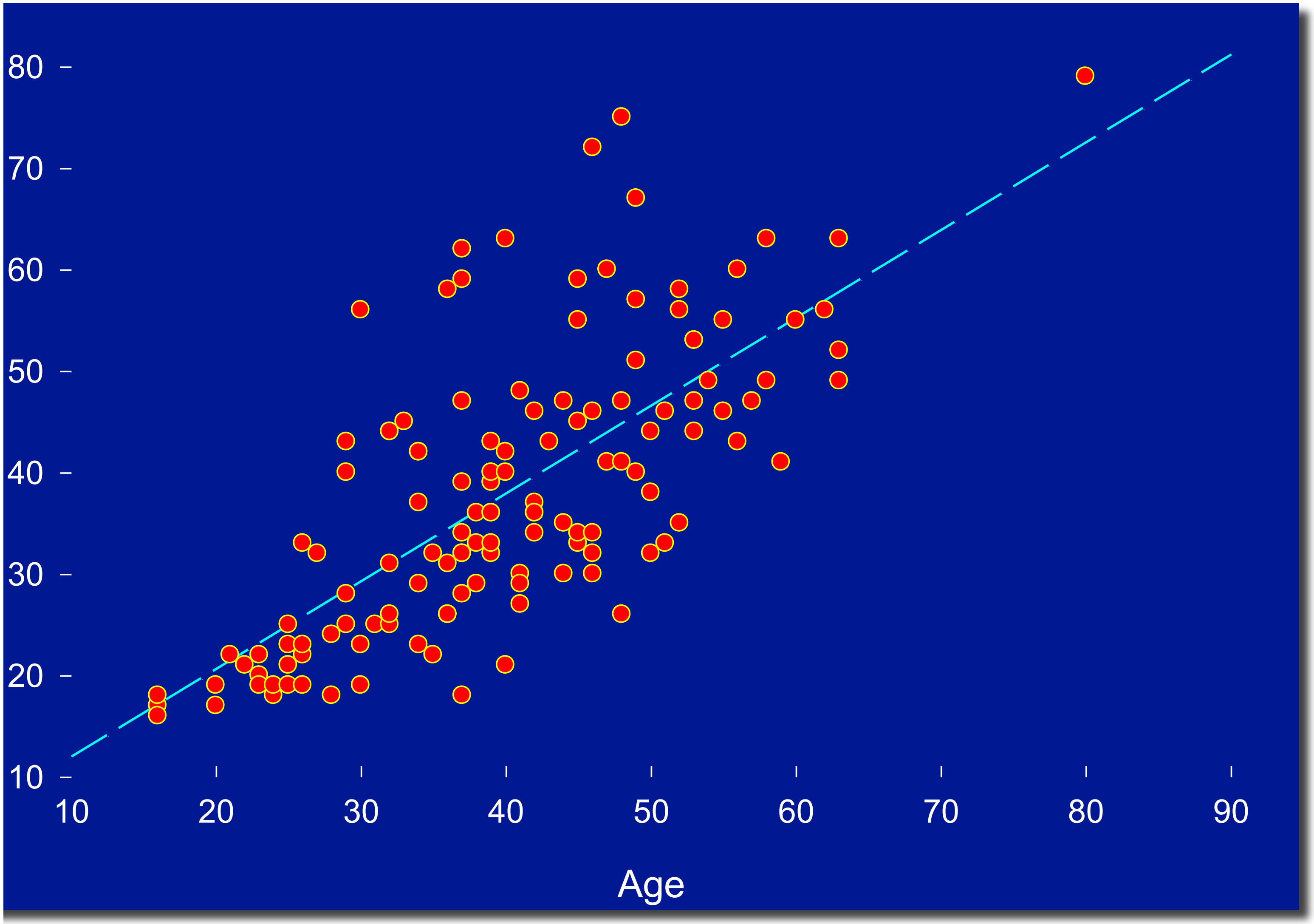
- Measuring relationships
  - Covariance
  - Pearson's correlation coefficient
- Nonparametric measures
  - Spearman's rho
  - Kendall's tau
- Interpreting correlations
  - Causality
- Partial correlations

What is a Correlation?

- It is a way of measuring the extent to which two variables are related.
- It does this by considering datapoints that have values along both variables

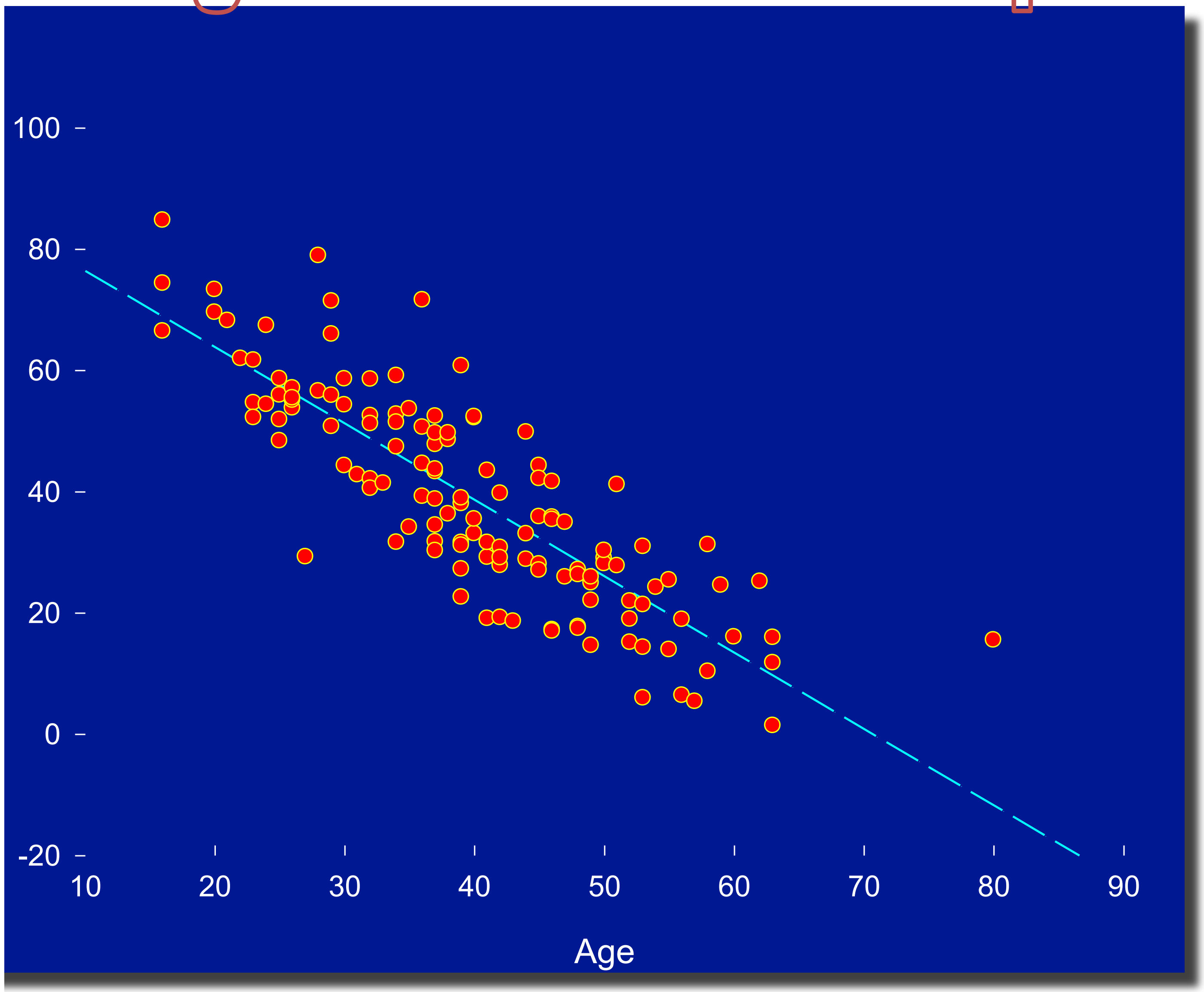
Positive Relationship

Age of
Best
Friend

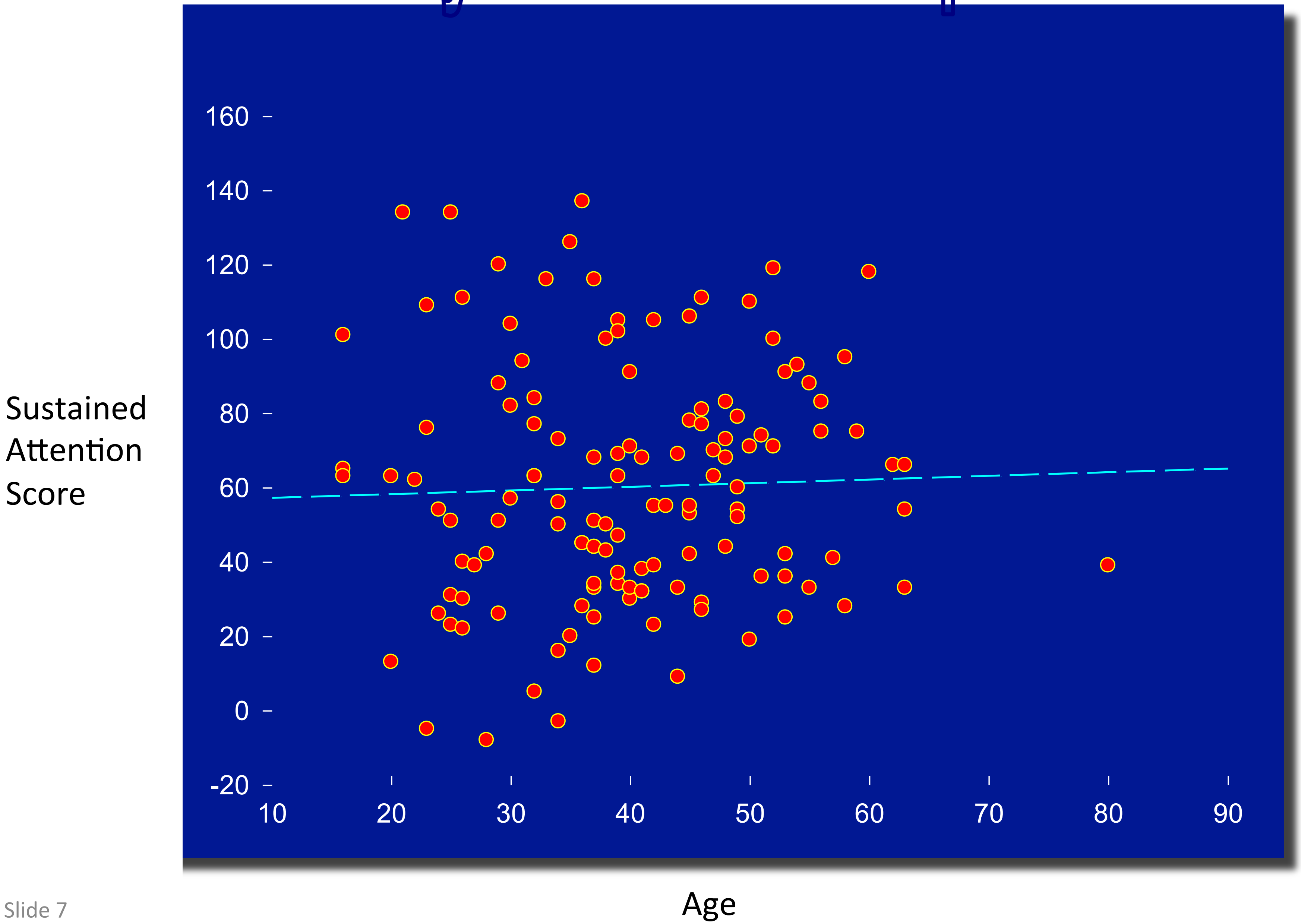


Negative Relationship

Memory
Recall
Task



Very Small Relationship



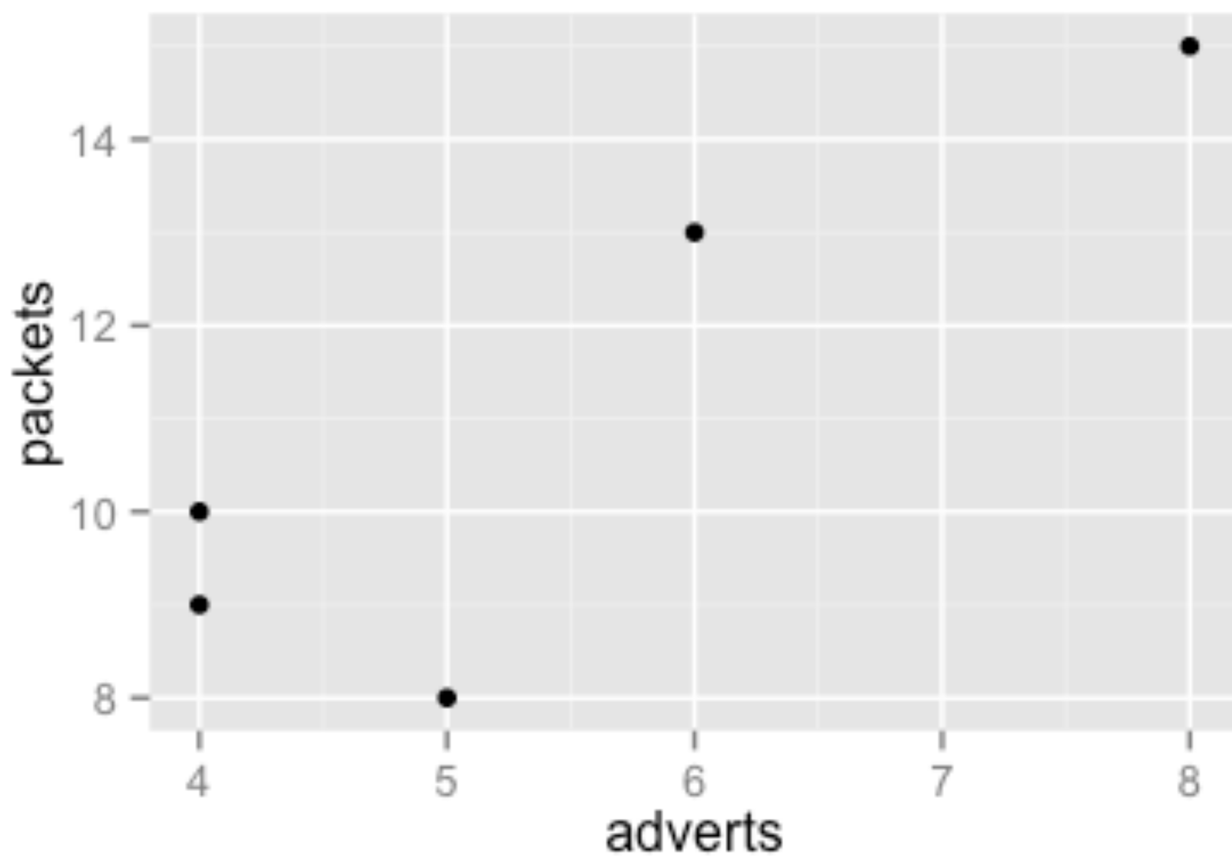
Measuring Relationships

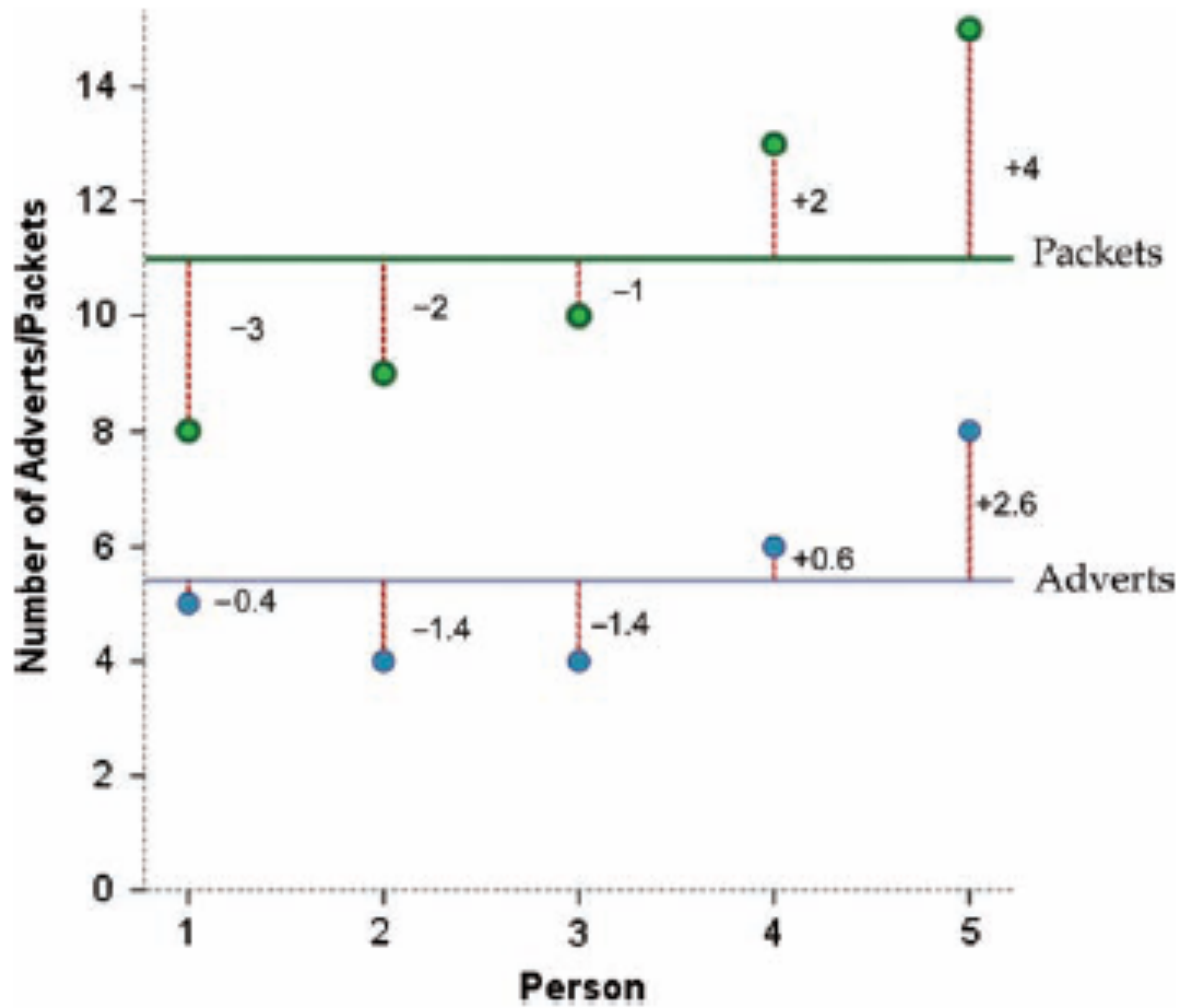
- We need to see whether as one variable increases, the other increases, decreases or stays the same.
- This can be done by calculating the covariance.
  - We look at how much each score deviates from the means of the two variables
  - Do the individual scores exhibit similar patterns in how they deviate from each variable?

- Observational Data
  - 5 participants
  - Number of television advertisements watched about a particular candy
  - Number of packets of candy bought

Participant:	1	2	3	4	5	Mean	S
Adverts Watched	5	4	4	6	8	5.4	1.67
Packets Bought	8	9	10	13	15	11.0	2.92

Scatterplot





Number of Adverts/Packets
Person
Packets
Adverts
−3
−2
−1
+2
+4
−0.4
−1.4
−1.4
+0.6
+2.6
0
2
4
6
8
10
12
14
1
2
3
4
5

Review of Variance

- The variance tells us by how much scores deviate from the mean for a single variable.

$$\text{variance } (s^2) = \frac{\text{SS}}{N-1} = \frac{\sum(x_i - \overline{x})^2}{N-1}$$

- Covariance is similar – we calculate an average of products of deviations

Covariance

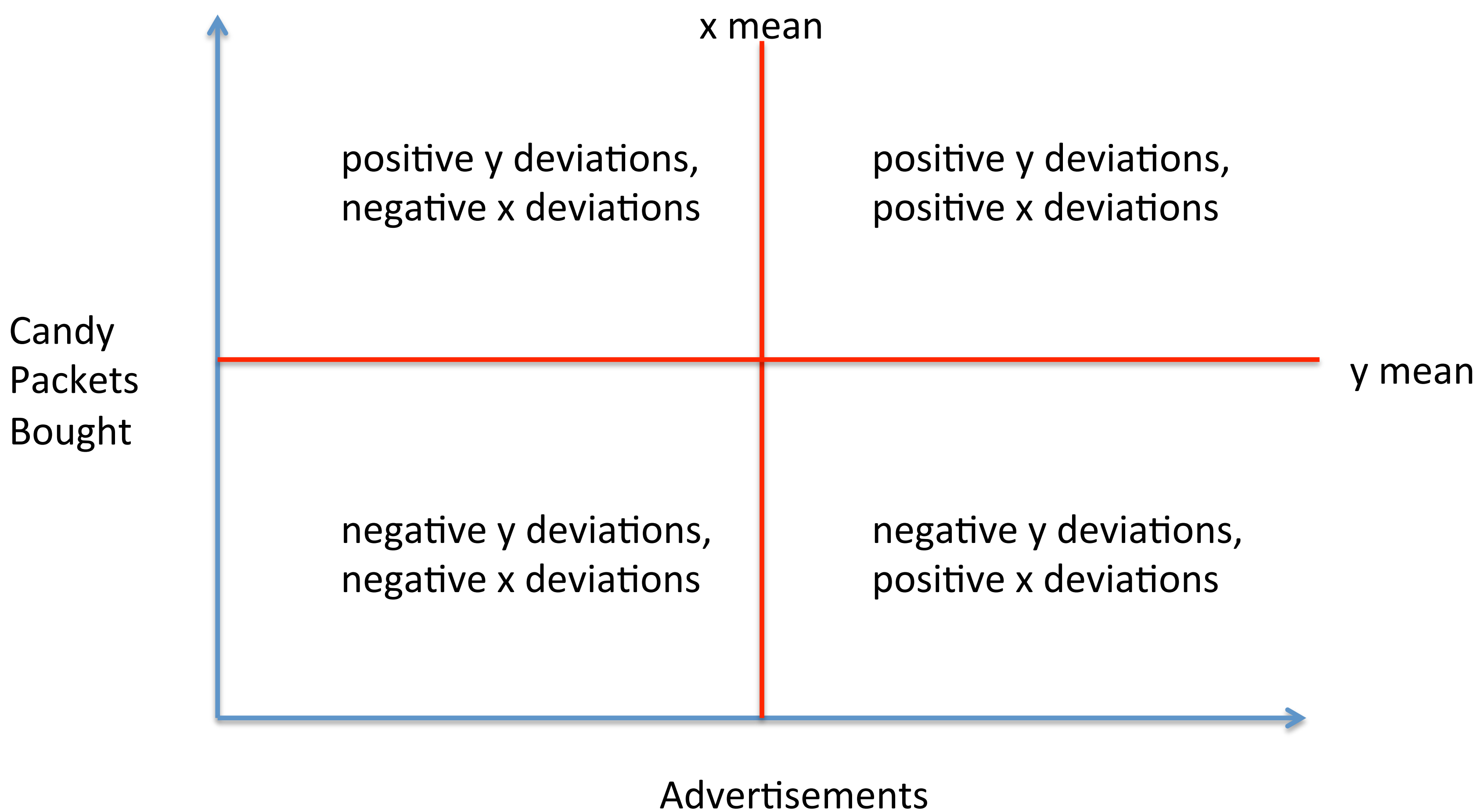
- Calculate the deviations between the mean and each score for the first variable (*x*).
- Calculate the deviations between the mean and each score for the second variable (*y*).
- In a pairwise fashion, multiply these deviations (to get the "cross-product deviations")
- Sum them
- The covariance is the average of the sum of cross-product deviations:

$$\operatorname{cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

Participant:	1	2	3	4	5	Mean	S
Adverts Watched	5	4	4	6	8	5.4	1.67
Packets Bought	8	9	10	13	15	11.0	2.92

$$
\begin{aligned}
\mathrm{cov}(x,y) &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1} \\
&= \frac{(-0.4)(-3) + (-1.4)(-2) + (-1.4)(-1) + (0.6)(2) + (2.6)(4)}{4} \\
&= \frac{1.2 + 2.8 + 1.4 + 1.2 + 10.4}{4} \\
&= \frac{17}{4} \\
&= 4.25
\end{aligned}
$$

Positive and Negative Cross-Product Deviations



Covariance Examples in RStudio

Problems with Covariance

- It depends upon the units of measurement.
  - E.g. the covariance of two variables measured in miles might be 4.25, but if the same scores are converted to kilometres, the covariance is 11.
- One solution: standardize it!
  - Divide by the standard deviations of both variables.
- The standardized version of covariance is known as Pearson's r or Pearson correlation coefficient

Things to Know about the Pearson correlation coefficient

- It varies between -1 and +1
  - 0 = no relationship
  - -1 = perfect negative correlation
  - 1 = perfect positive correlation
- It is an effect size
  - ±.1 = small effect
  - ±.3 = medium effect
  - ±.5 = large effect

The Correlation Coefficient

$$r = \frac{\mathrm{cov}_{xy}}{s_x s_y}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(N-1) s_x s_y}$$

Why can't r be greater than 1 or less than -1?

$$r = \frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{n - 1 * \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} * \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}}$$

$$r = \frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{n - 1 * \frac{\sqrt{\sum(x_i - \bar{x})^2}}{\sqrt{n-1}} * \frac{\sqrt{\sum(y_i - \bar{y})^2}}{\sqrt{n-1}}}$$

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

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

Cauchy-Schwarz Inequality

Coefficient of Determination R^2

- Just r to the power of 2 (then multiply by 100)!
- measures (in percentage) amount of variance of one variable accounted for by the other
  - symmetrical

Correlations in RStudio

Reporting the Results

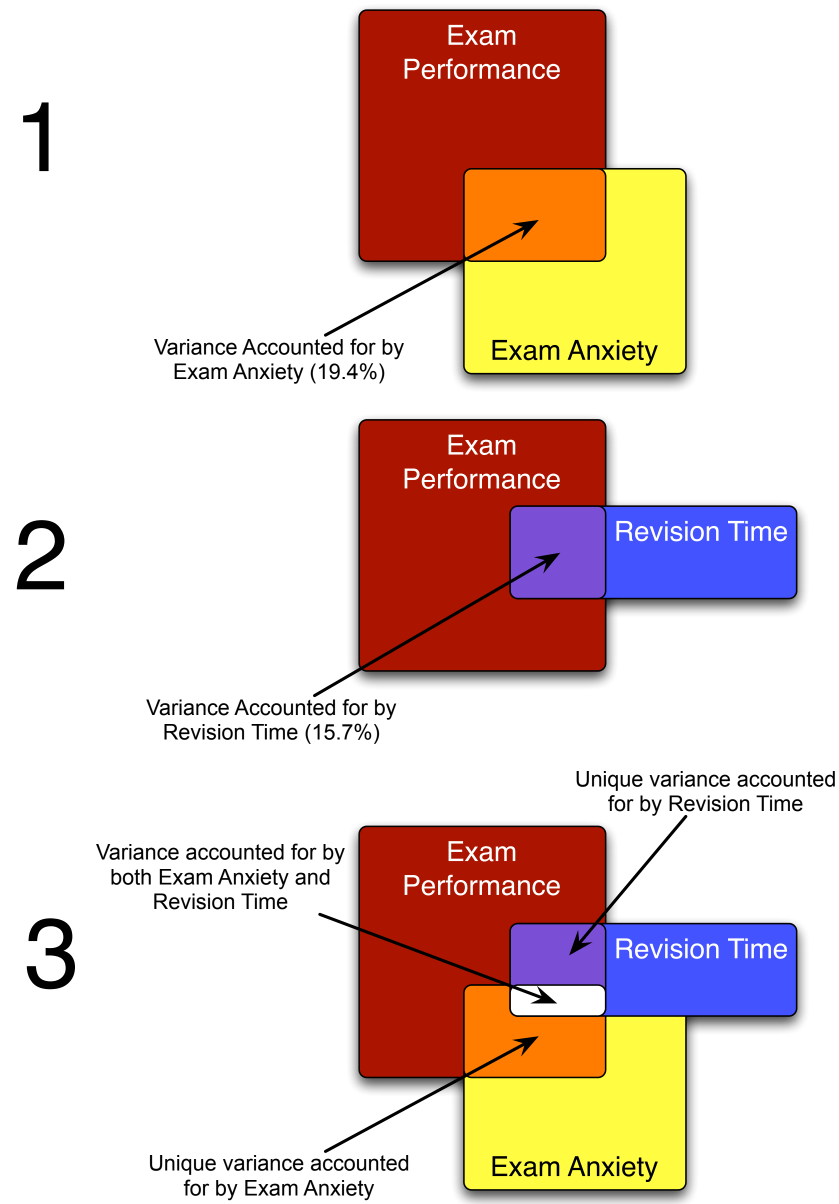
- "Exam performance was significantly correlated with exam anxiety, r = -.44, and time spent revising, r = .40; the time spent revising was also correlated with exam anxiety, r = -.71 (all ps < .001)."
- Mention if you used a one-tailed test

Correlation and Causality

- The third-variable problem:
  - there may be other measured or unmeasured variables affecting the results.
- Direction of causality:
  - Correlation coefficients say nothing about which variable causes the other to change.
  - In observational studies, you don't know about temporal order in which events occurred

Bivariate vs. Partial Correlations

- Bivariate
  - Correlation between 2 variables
- Partial
  - Measure the relationship between two variables, controlling for the effect that a third variable has on them both.
    - R^2 of Exam Performance and Exam Anxiety
      - » $-.44^2 \times 100 = 19.4$
    - R^2 of Exam Performance and Revision Time
      - » $.397^2 \times 100 = 15.7$
    - R2 of Revision Time and Exam Anxiety
      - » $.-.709^2 \times 100 = 50.2$

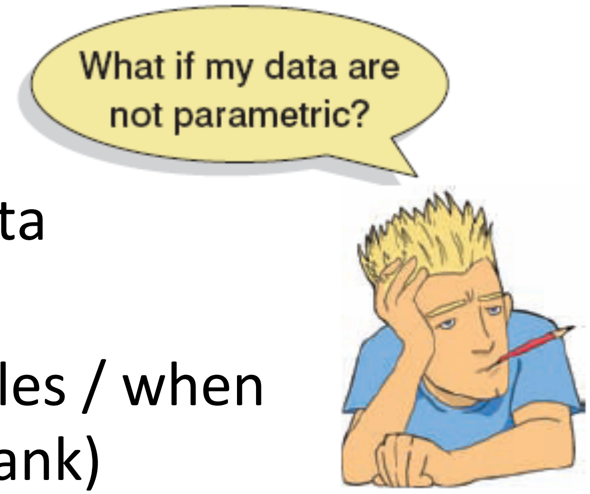

1
Exam Performance
Exam Anxiety
Variance Accounted for by Exam Anxiety (19.4%)
2
Exam Performance
Revision Time
Variance Accounted for by Revision Time (15.7%)
3
Unique variance accounted for by Revision Time
Variance accounted for by both Exam Anxiety and Revision Time
Exam Performance
Revision Time
Exam Anxiety
Unique variance accounted for by Exam Anxiety

Partial Correlation in RStudio

Assumptions of Pearson's R

- Interval/ratio data
- Significance test assumes normal sampling distribution
  - assume it is if sample data is normal
  - or if sample size is large

Non-parametric Correlation



- Spearman's rho
  - Pearson's correlation on the ranked data
- Kendall's tau
  - Better than Spearman's for small samples / when there are a lot of ties (values with same rank)
- World's Biggest Liar competition
  - 68 contestants
  - Measures
    - Where they were placed in the competition (first, second, third, etc.)
    - Creativity questionnaire (maximum score 60)
    - Alternative hypothesis: more creative people should be better liars (i.e., place higher in competition)

Calculating Spearman's Rho

- Same as Pearson's r!

$$r_s = \frac{\text{cov}_{xy}}{s_x s_y}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(N-1) s_x s_y}$$

- BUT your data is ordinal (naturally or you generate a rank order)
  – (3.3, 9.2, -1.4, 5.5) -> (2, 4, 1, 3)

Spearman's Rho

- Checking correlation in R
  - cor(liarData$Position, liarData$Creativity, method = "spearman")

    [1] -0.3732184

- Checking significance of correlation
  - cor.test(liarData$Position, liarData$Creativity, alternative = "less", method = "spearman")

- Note you can calculate R^2 for rho
  - proportion in the variance of the ranks that the two variables share

Spearman's Rho Output

Spearman's rank correlation rho

data: liarData$Position and liarData$Creativity

S = 71948.4, p-value = 0.0008602

alternative hypothesis: true rho is less than 0

sample estimates:

rho

-0.3732184

Calculating Kendall's Tau

- tau = (C – D) / (C + D)
  - C = # concordant pairs
  - D = # discordant pairs
- Best animals

	Hippo	Lion	Koala	Zebra	Panda	TOTAL
Person 1	1	2	3	4	5	
Person 2	2	4	1	5	3	
C	3	1	2	0		6
D	1	2	0	1		4

tau = (6-4)/(6+4) = .2

Kendall’s Tau

- To carry out Kendall’s correlation on the World’s Biggest Liar data simply follow the same steps as for Pearson and Spearman correlations but use *method = “kendall”*:

```
cor(liarData$Position, liarData$Creativity, method = "kendall")

cor.test(liarData$Position, liarData$Creativity, alternative = "less", method = "kendall")
```

Kendall's Tau

- The output is much the same as for Spearman's correlation.

```
    Kendall's rank correlation tau
data:  liarData$Position and liarData$Creativity
z = -3.2252, p-value = 0.0006294
alternative hypothesis: true tau is less than 0
sample estimates:
     tau
-0.3002413
```