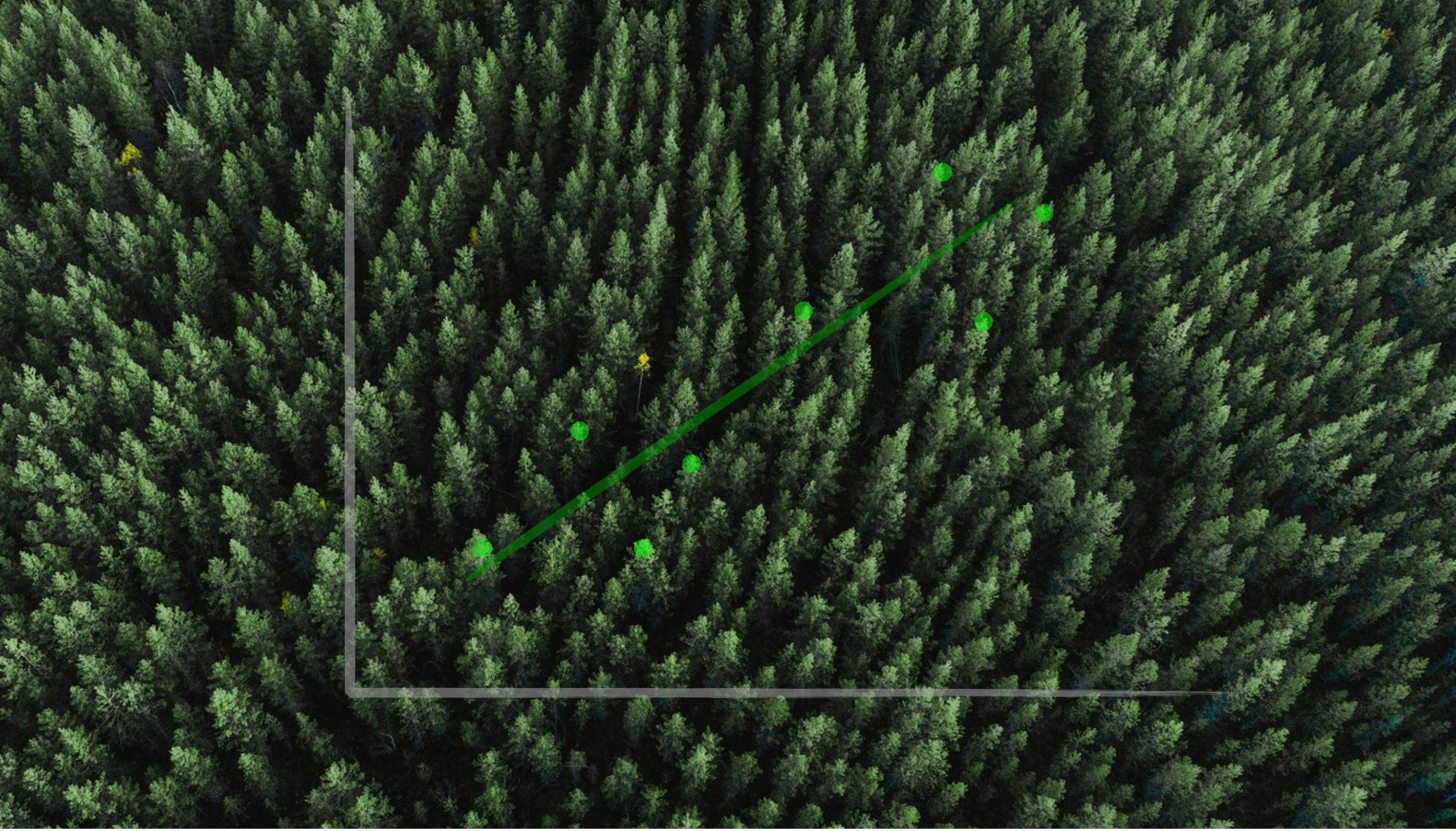


Correlation analysis: Looking at relationships in our data

Methods 1, E2021 - Lecture 5
Tuesday 5/10/2020
Fabio Trecca

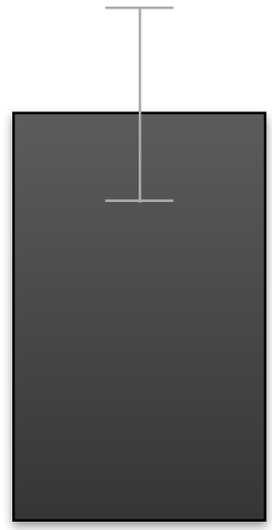
Attendance registration

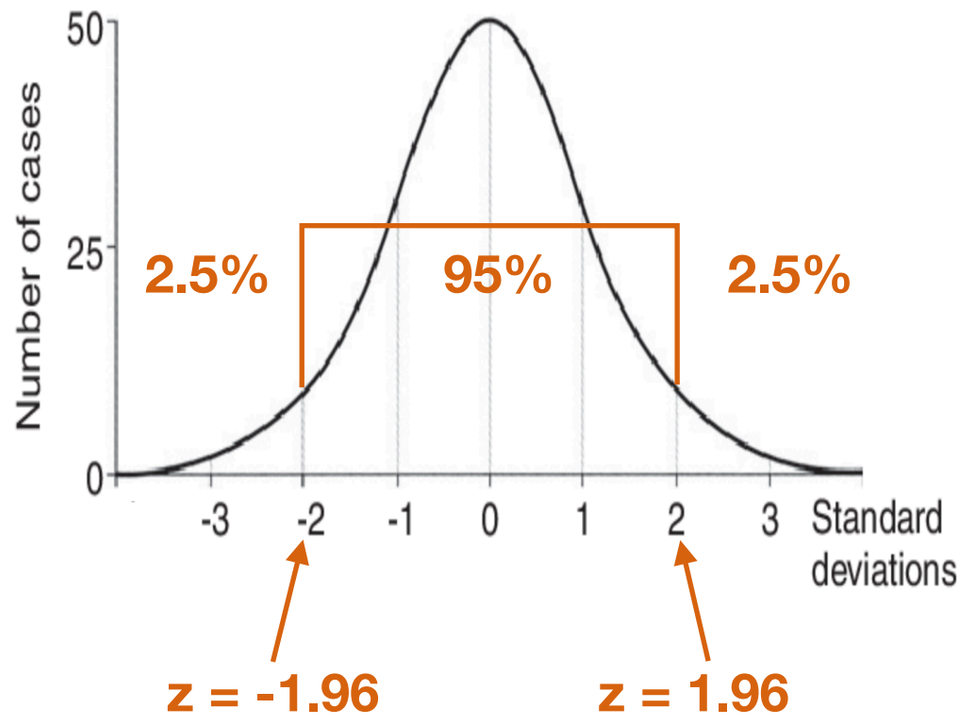
Check in using the PIN-code


QUIZ
TIME

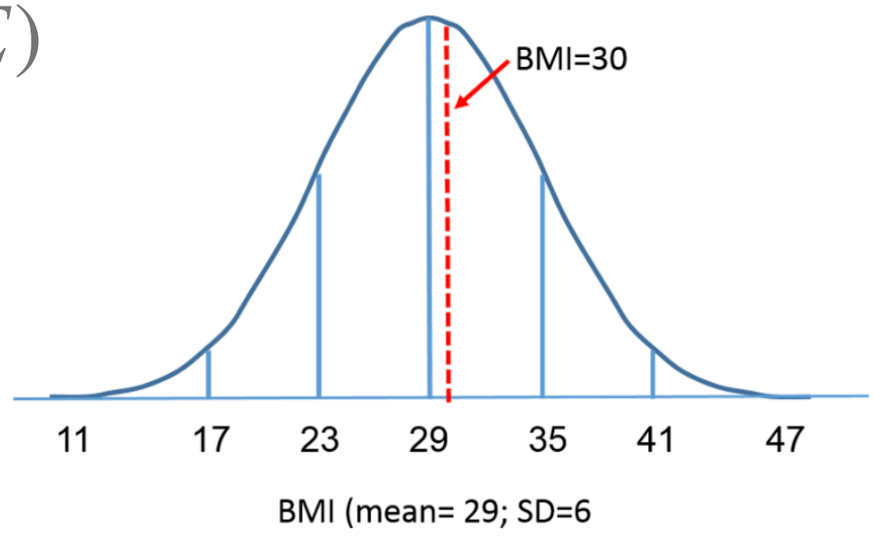
Quiz time (1)

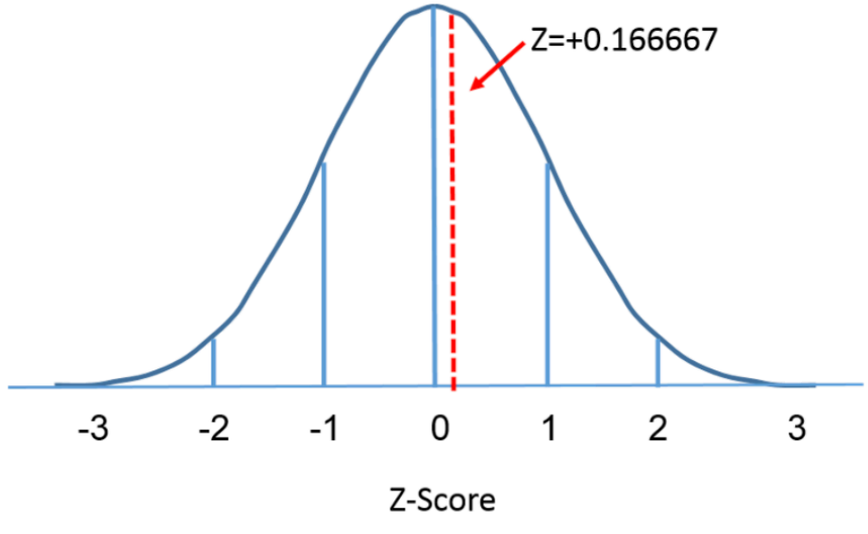
- $z_i = \frac{x_i - \bar{x}}{s}$
- $\bar{x} \pm z \times \frac{\sigma}{\sqrt{n}}$



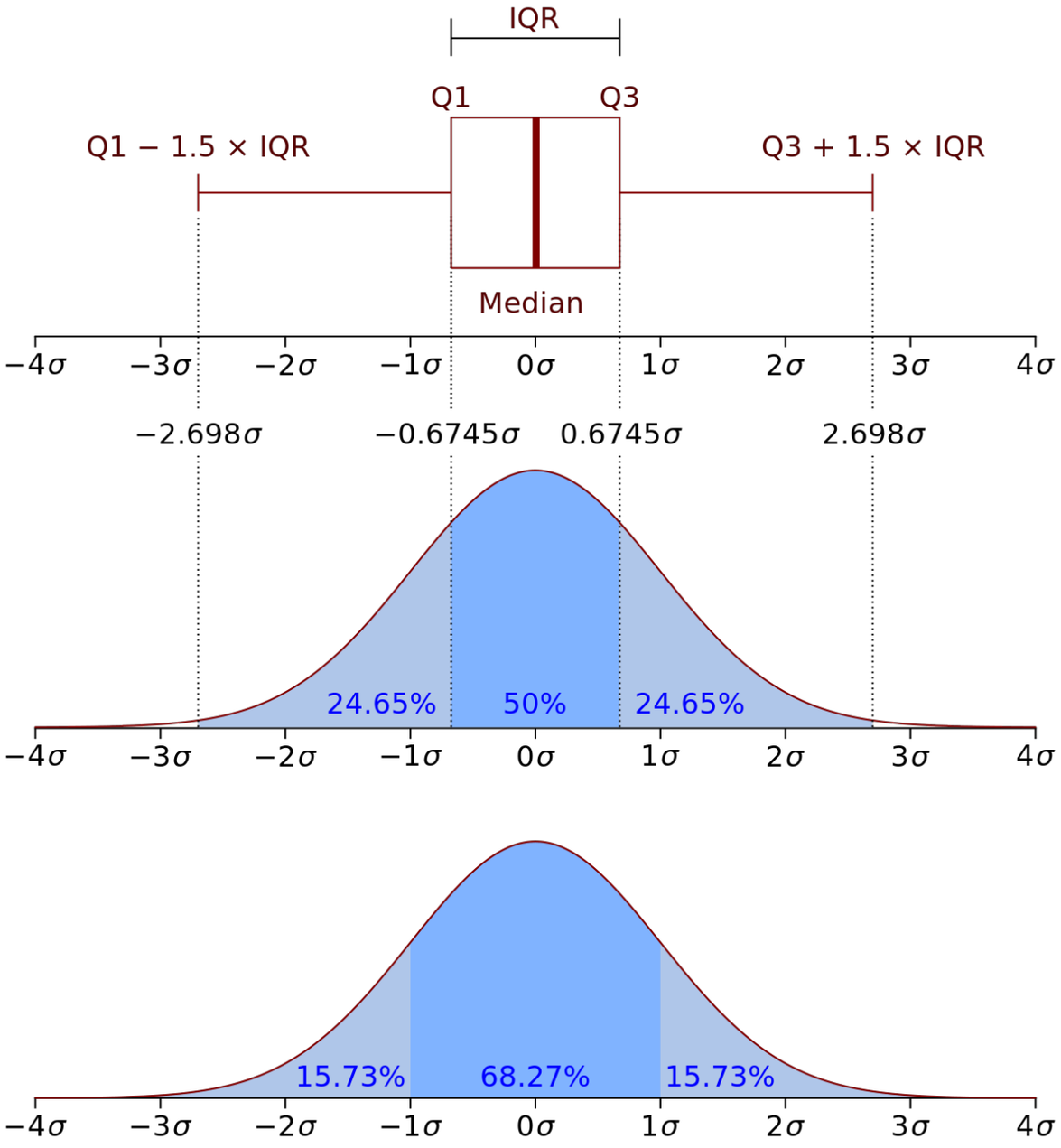


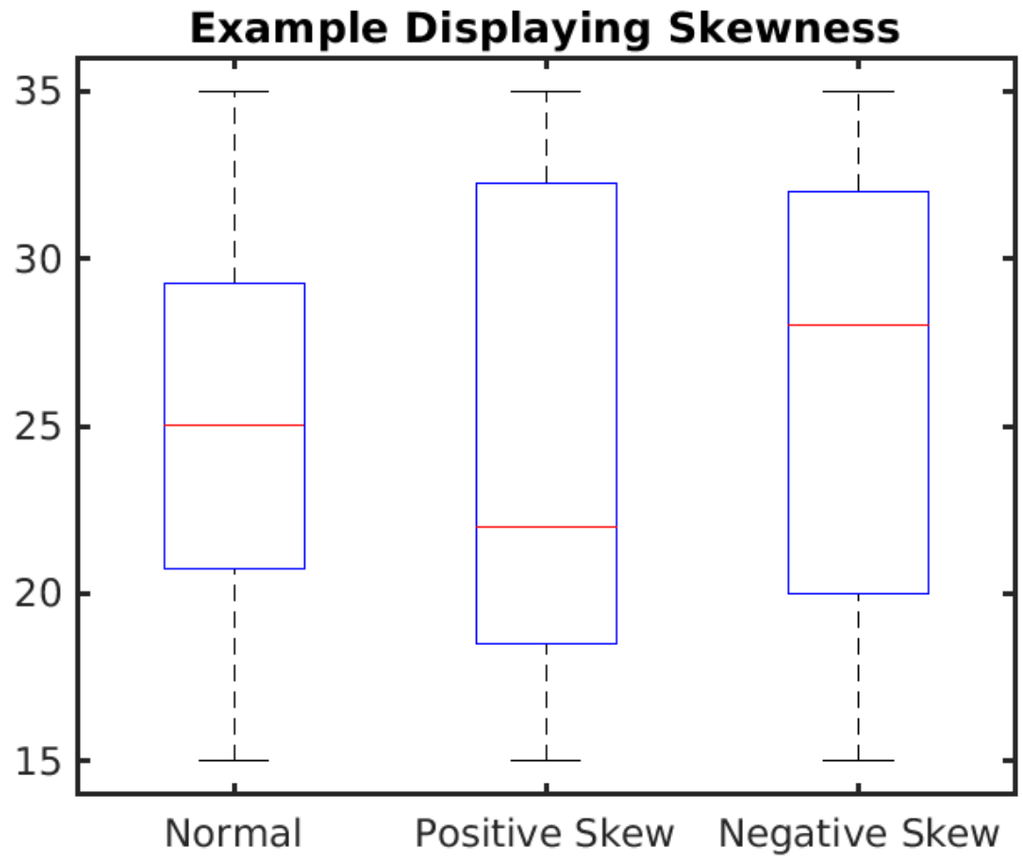
- $\bar{x} + (1.96 \times SE)$
- $\bar{x} - (1.96 \times SE)$





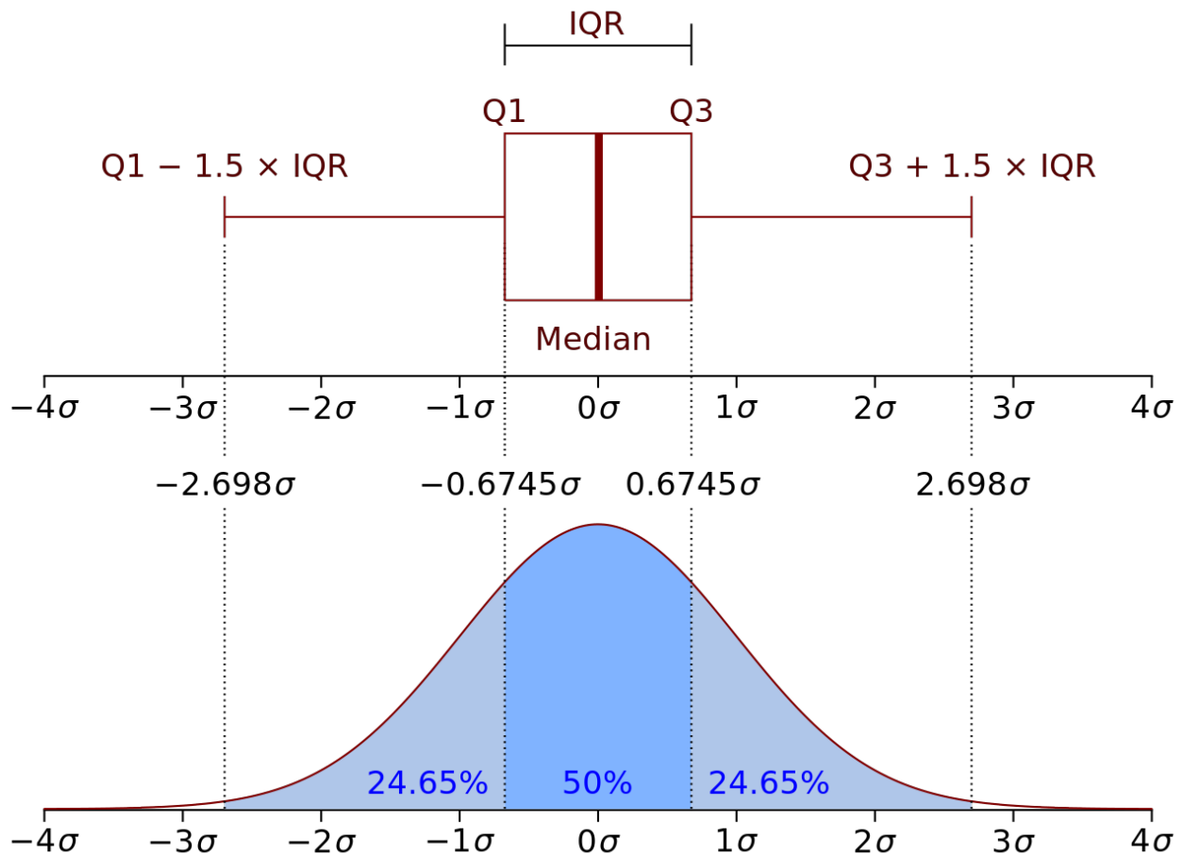
Quiz time (2)


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%


Example Displaying Skewness
35
30
25
20
15
Normal
Positive Skew
Negative Skew

Interquartile Range

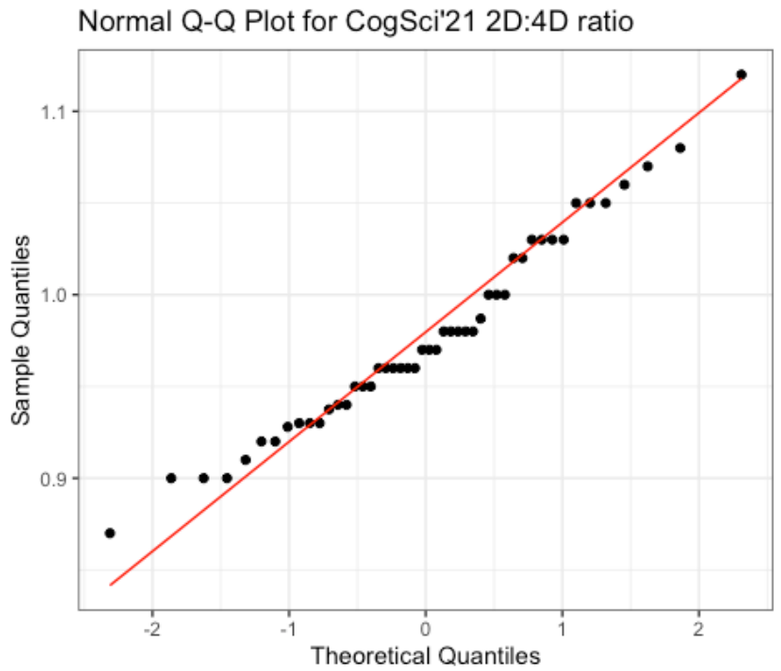
- A measure of statistical dispersion equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles

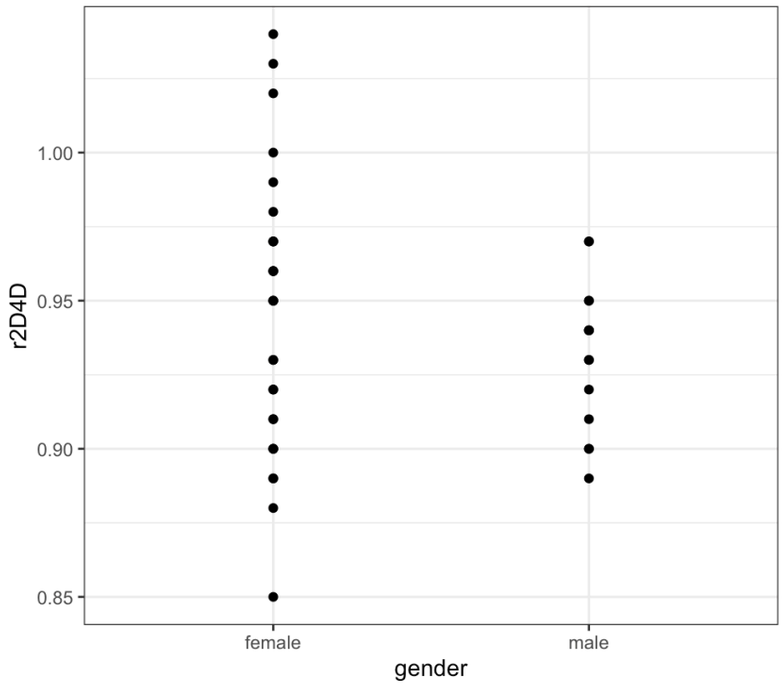


- IQR = Q3 – Q1
- Quantiles
- Quartiles
- Percentiles

Quiz time (3)

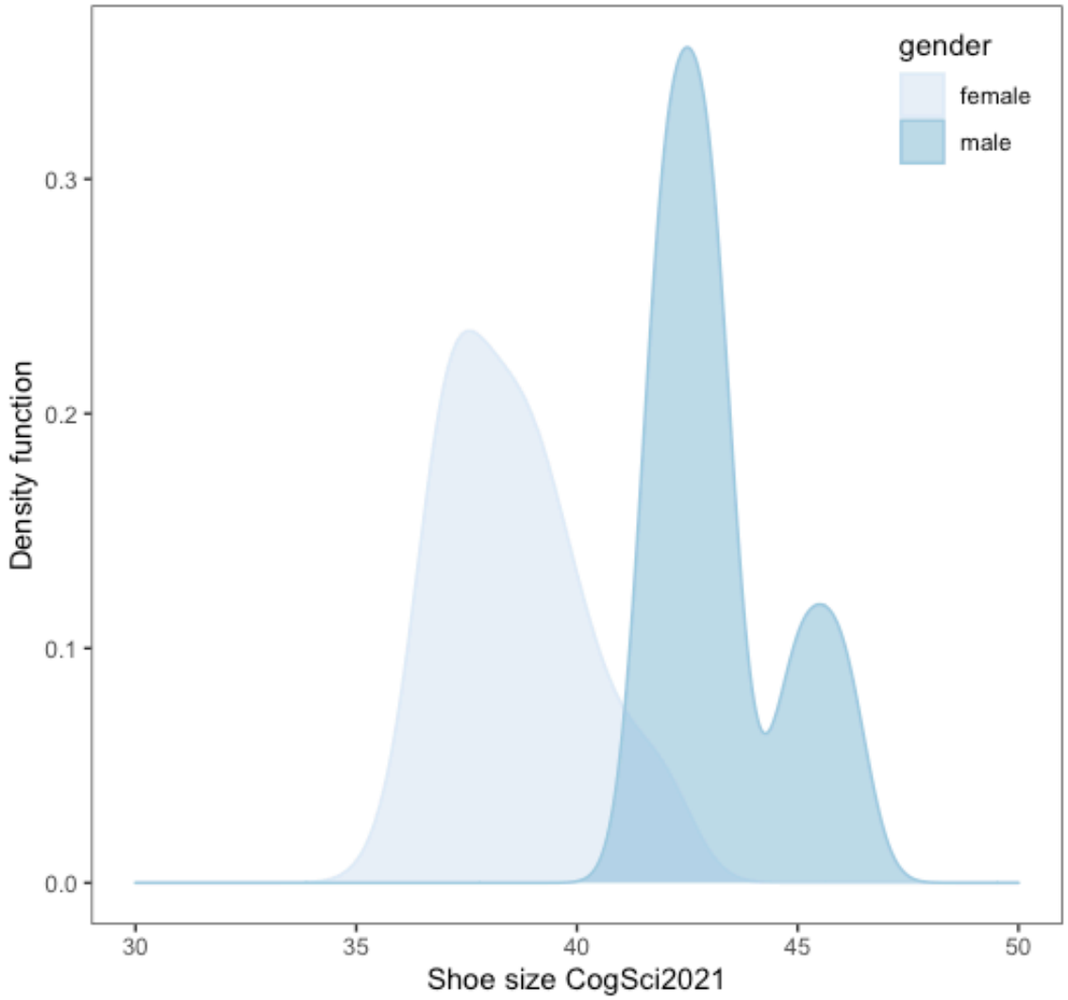
- Assessing normality:
  - skewness/kurtosis
  - Shapiro-Wilk test
  - `pastecs::stat.desc`
- Assessing homoscedasticity:
  - Levene's test
  - `car::leveneTest`





Descriptive vs Inferential statistical tests (1)

- Descriptive statistical tests
  - summary characterization of observed data
  - only limited to them, not generalizable
  - expressed in descriptive statistics (mean, median, SD, kurtosis, …)
- Inferential statistical tests
  - quantify whether observed effects are generalizable to unobserved data (population parameters)
  - expressed in test statistics (SE, r, t, z, p, effect size)



Descriptive vs Inferential statistical tests (2)

- Correlation:
  tests the relationship between two continuous variables

← today

- T-test:
  tests whether the means of 2 groups differ significantly

- ANOVA:
  tests whether the means of 3+ groups differ significantly

- Linear regression:
  tests whether 1+ continuous/categorical independent variables predict a continuous dependent variable

- Logistic regression:
  tests whether 1+ continuous/categorical independent variables predict a categorical (binary) dependent variable

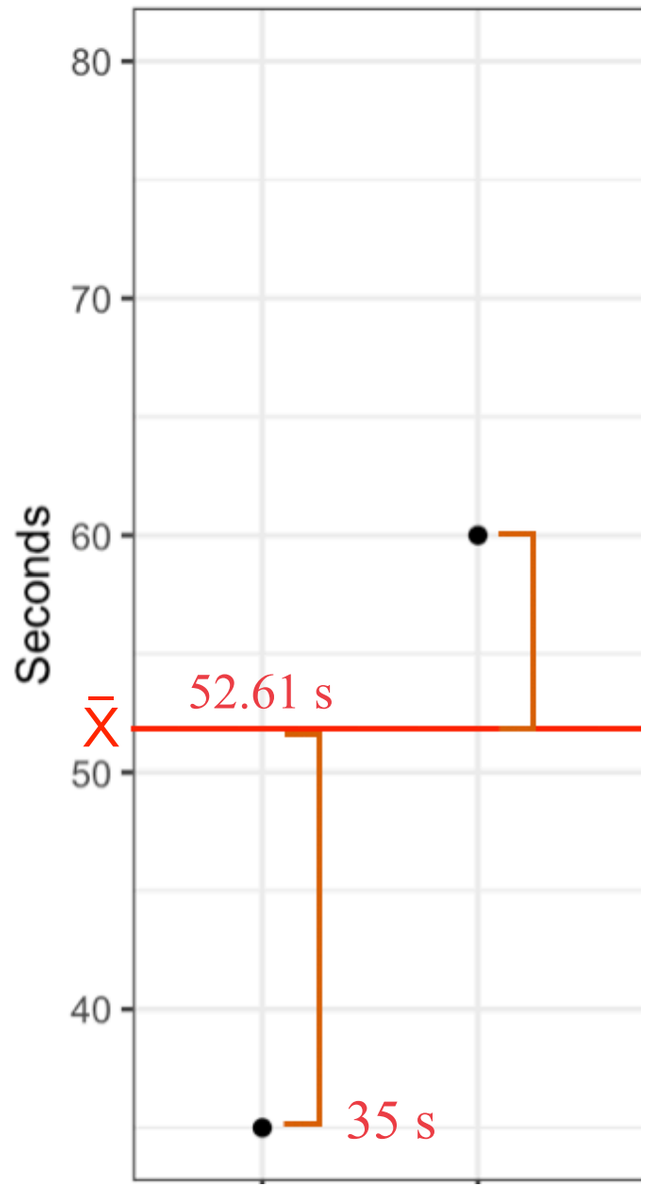
Remember variance (σ^2)?

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

- = mean deviance of each observed point from the basic model ($\bar{x}$)

- quantifies the variability of **one** measurement in a variable

- a way of quantifying whether $\bar{x}$ is a good model for the data

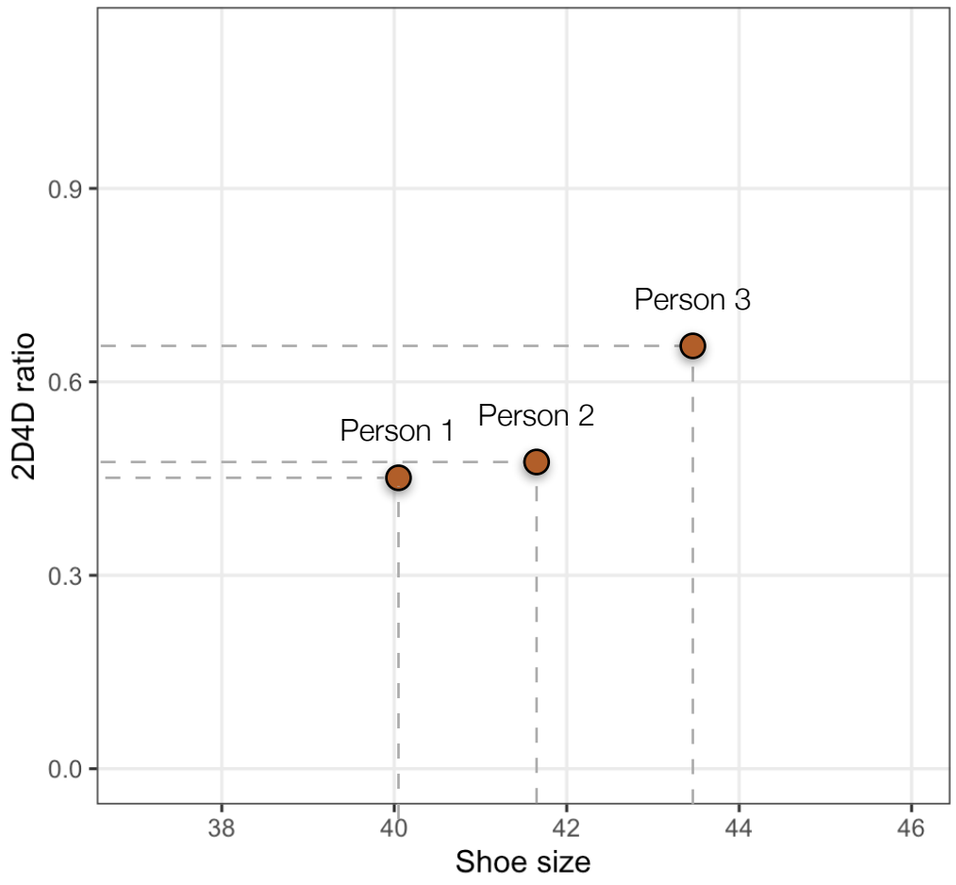


Covariance (1)

- A measure of how two measurements on two different variables vary together

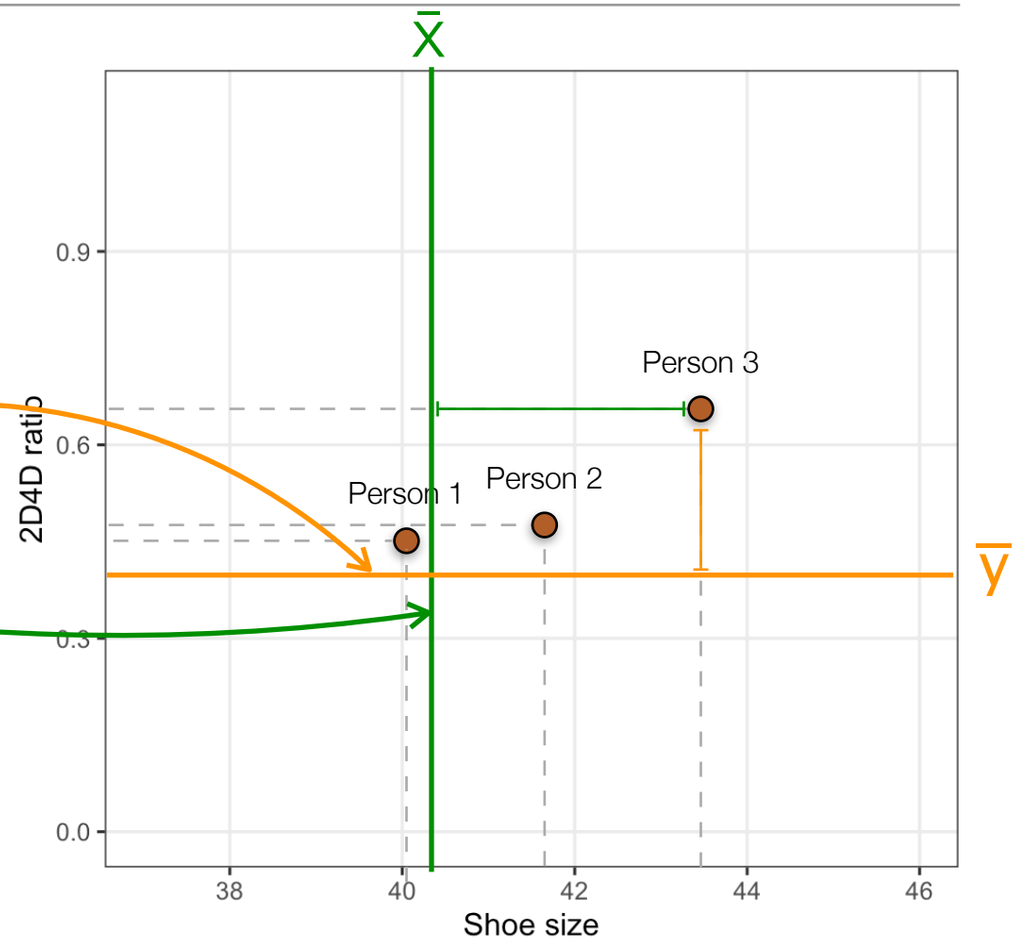
- $$cov(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

- *= average cross-product deviance*



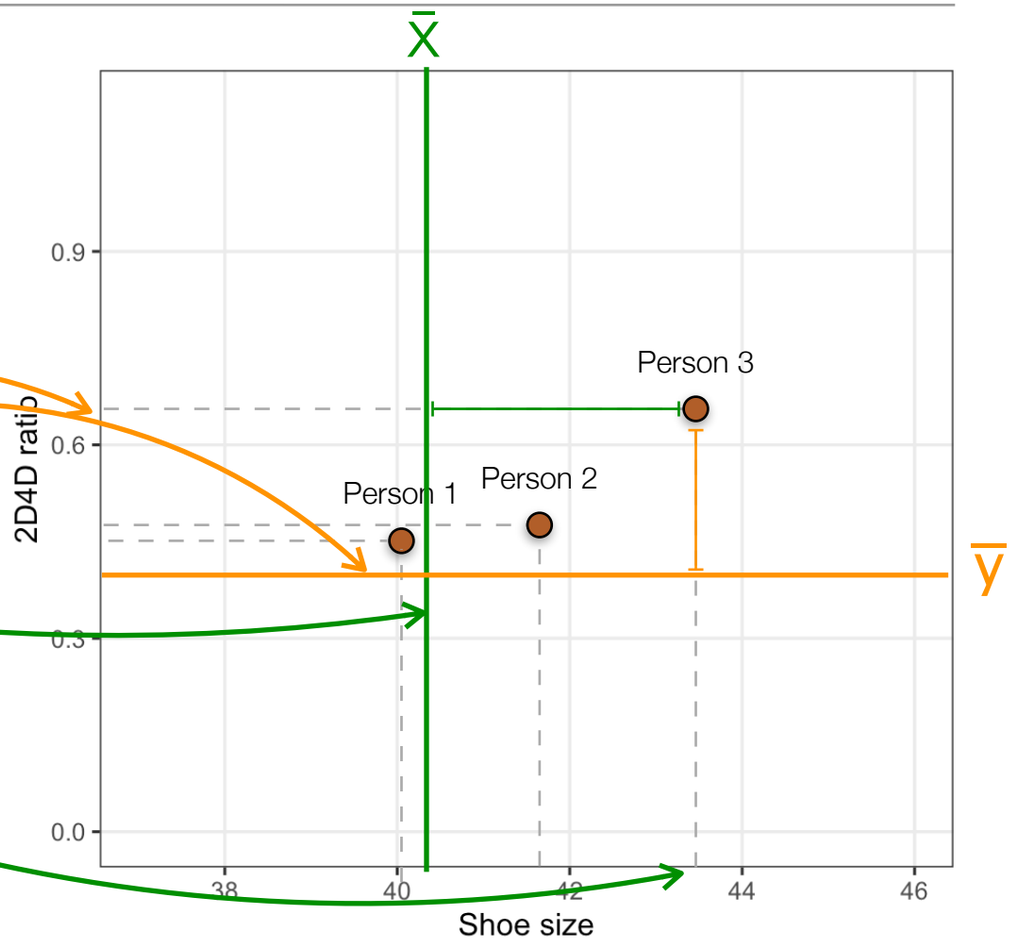
Covariance (1)

- A measure of how two measurements on two different variables vary together
- $$cov(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$
- = average cross-product deviance



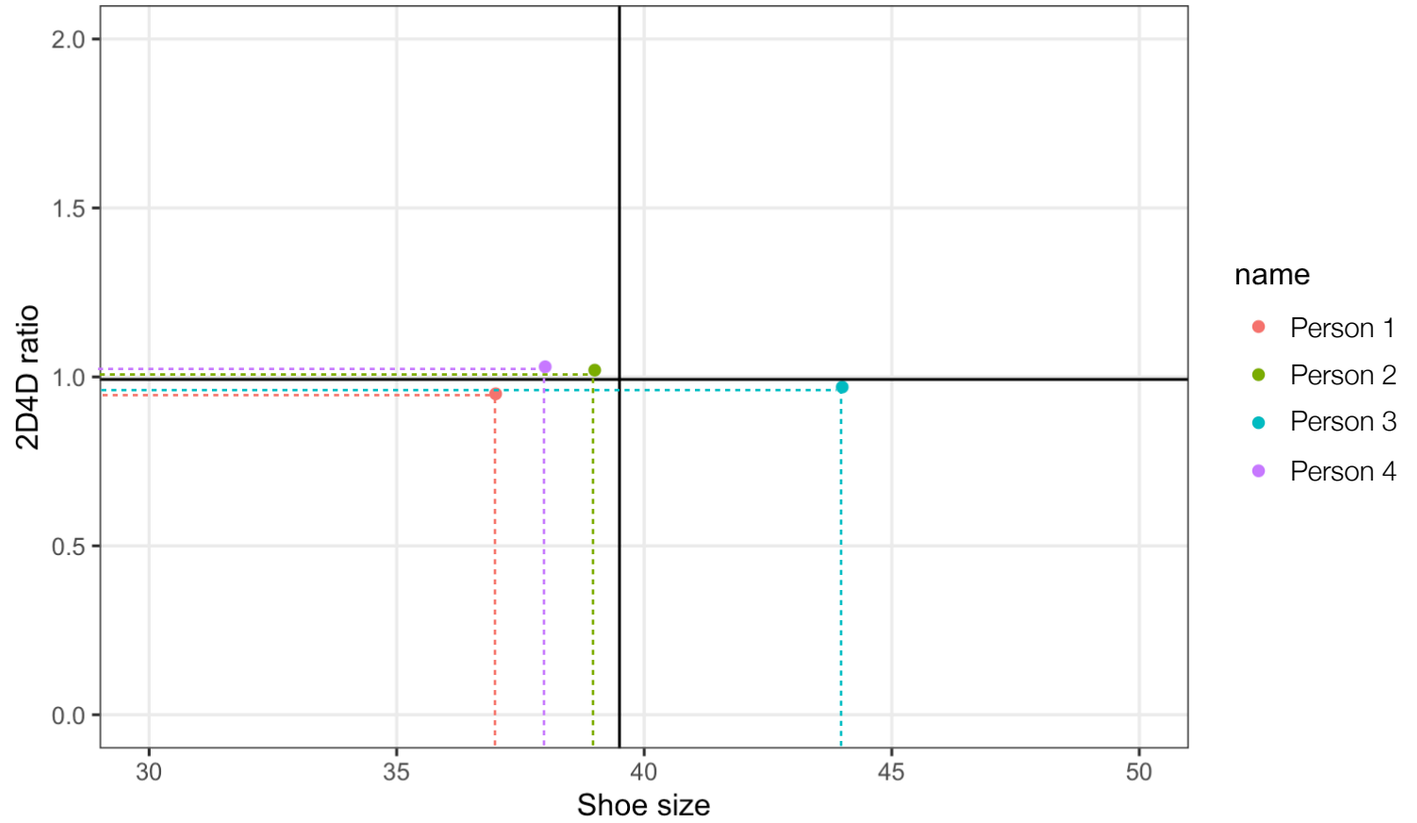
Covariance (1)

- A measure of how two measurements on two different variables vary together
- $$cov(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$
- = average cross-product deviance



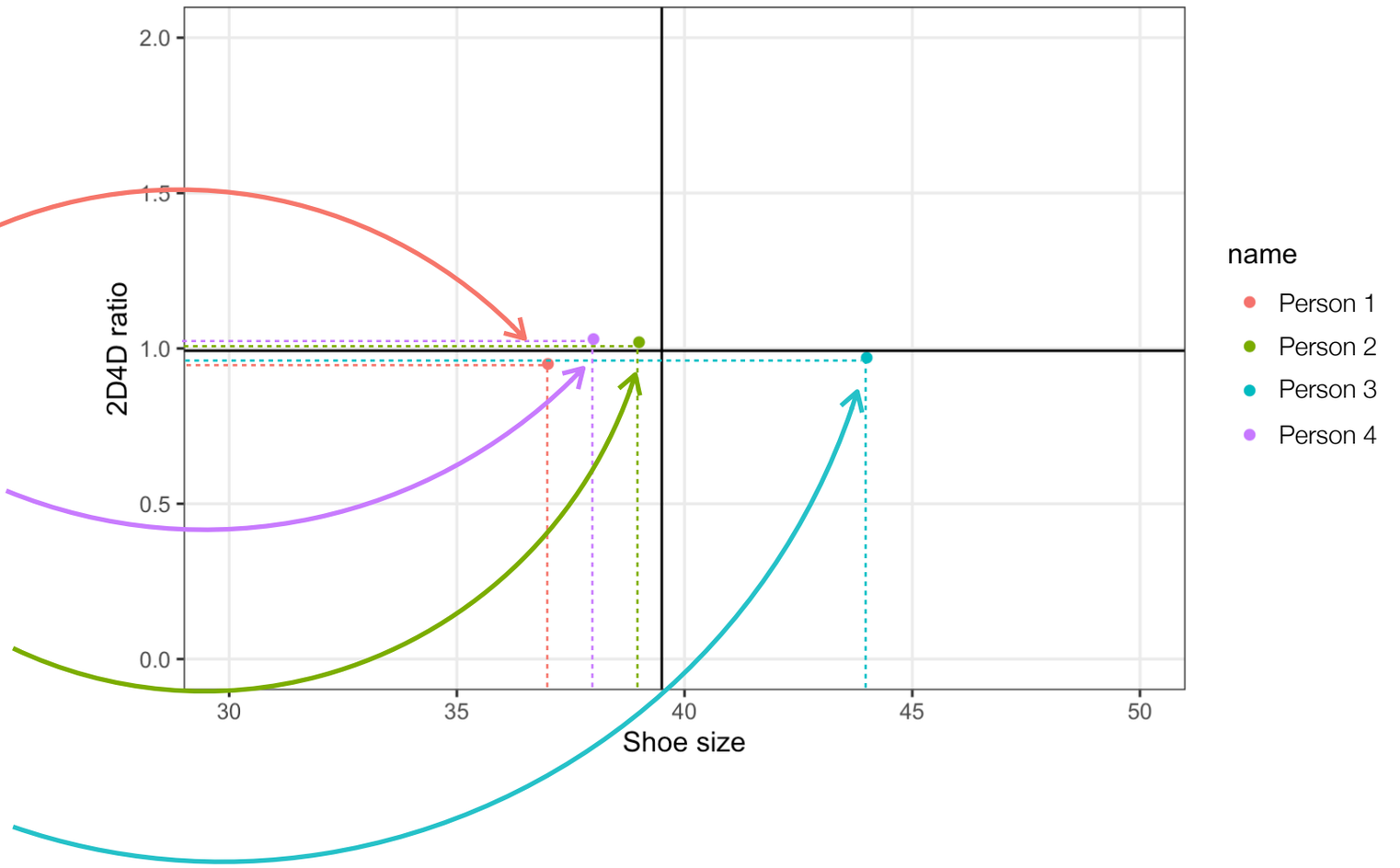
Covariance (2)

- $cov(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$
  - $(37 - 39.5)(0.95 - 0.99) = 0.1$
  - $(38 - 39.5)(1.03 - 0.99) = -0.06$
  - $(39 - 39.5)(1.02 - 0.99) = -0.015$
  - $(44 - 39.5)(0.97 - 0.99) = -0.09$
- $\frac{0.1 + (-0.06) + (-0.015) + (-0.09)}{4 - 1} =$
- $= -0.022$



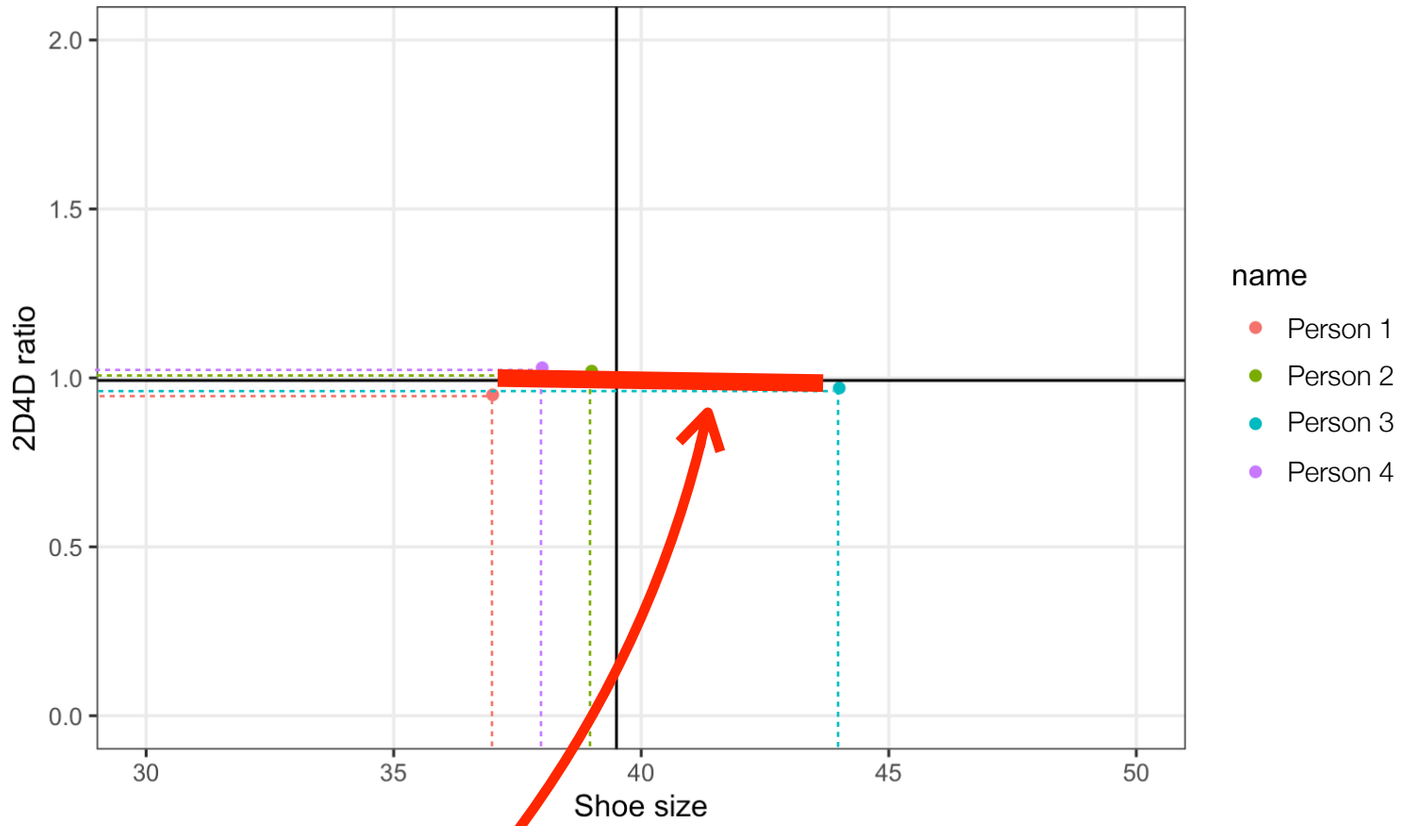
Covariance (2)

- $cov(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$
  - $(37 - 39.5)(0.95 - 0.99) = 0.1$
  - $(38 - 39.5)(1.03 - 0.99) = -0.06$
  - $(39 - 39.5)(1.02 - 0.99) = -0.015$
  - $(44 - 39.5)(0.97 - 0.99) = -0.09$
- $\frac{0.1 + (-0.06) + (-0.015) + (-0.09)}{4 - 1} =$
- $= -0.022$



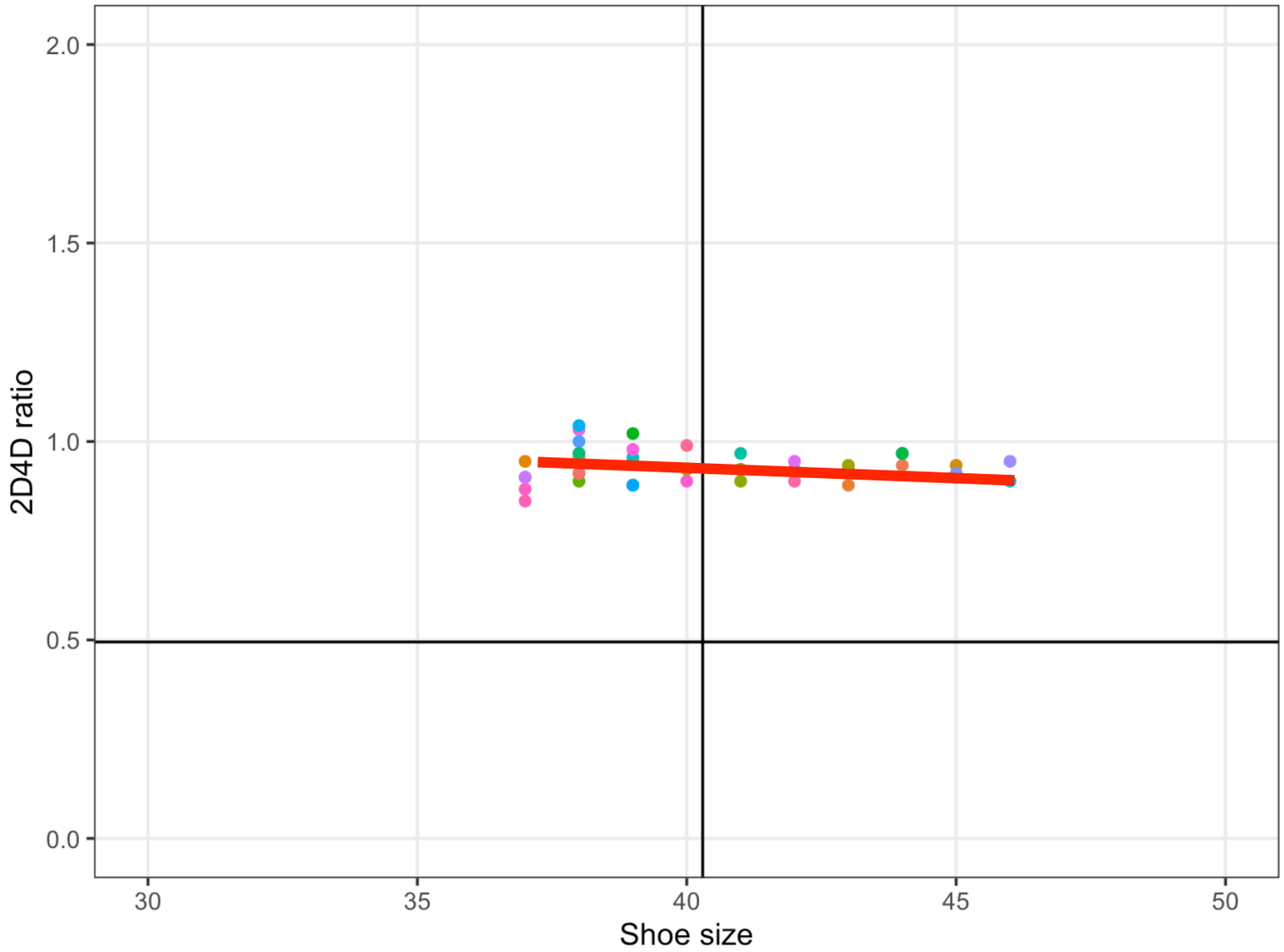
Covariance (2)

- $cov(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$
  - $(37 - 39.5)(0.95 - 0.99) = 0.1$
  - $(38 - 39.5)(1.03 - 0.99) = -0.06$
  - $(39 - 39.5)(1.02 - 0.99) = -0.015$
  - $(44 - 39.5)(0.97 - 0.99) = -0.09$
- $\frac{0.1 + (-0.06) + (-0.015) + (-0.09)}{4 - 1} =$
- $= -0.022$



Covariance (3)

- Positive covariance = positive relationship
- Negative covariance = negative relationship
- Covariance around zero = no relationship



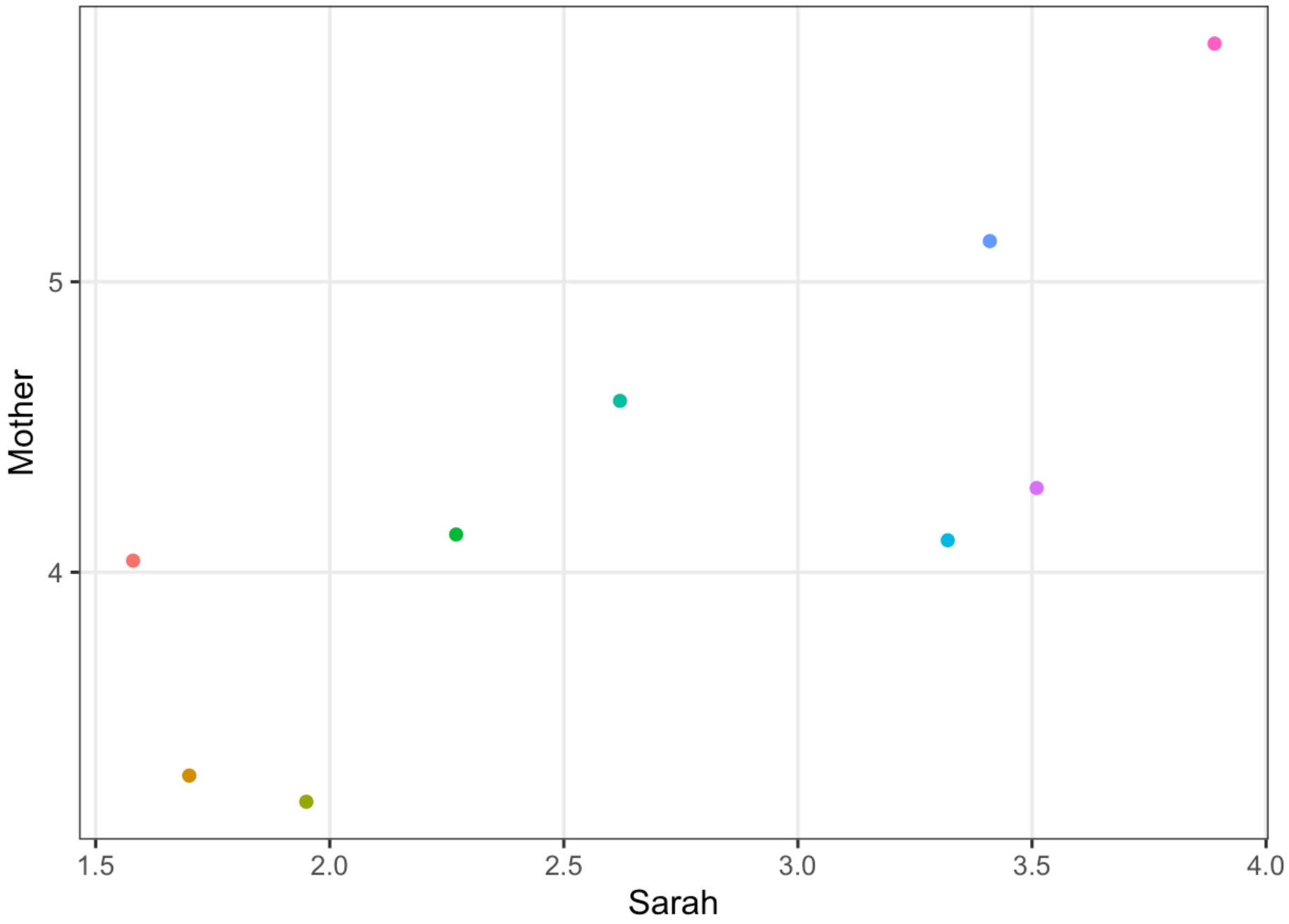
Exercise (1)

- Calculate the covariance coefficient for Sarah's and her mother's Mean Length of Utterances using the data here

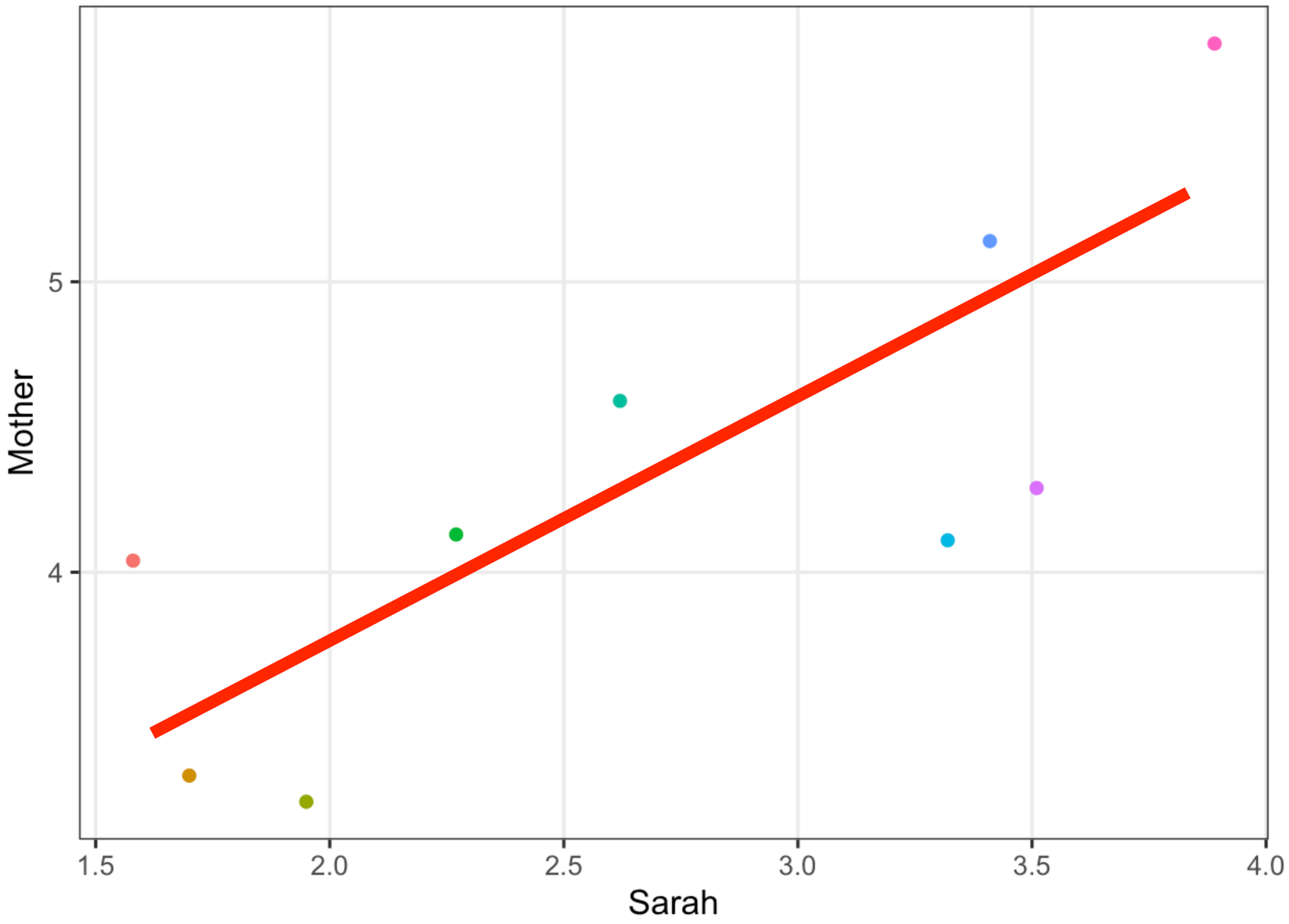
- *Mean Length of Utterance*: a measure of linguistic productivity in children, traditionally calculated by dividing the number of morphemes by the number of utterances

time	Sarah	Mother
06/1963	1.95	3.21
12/1963	1.58	4.04
02/1964	1.7	3.3
09/1964	2.27	4.13
12/1964	2.62	4.59
03/1965	3.32	4.11
06/1965	3.51	4.29
09/1965	3.89	5.82
12/1965	3.41	5.14

Exercise (2)



Exercise (2)



Exercise (3)

- sum(
  - (1.95 - mean(Sarah)) *
    (3.21 - mean(Mother)),
  - (1.58 - mean(Sarah)) *
    (4.04 - mean(Mother)),
  - …) /
  - (9 - 1) =
  - **0.554**

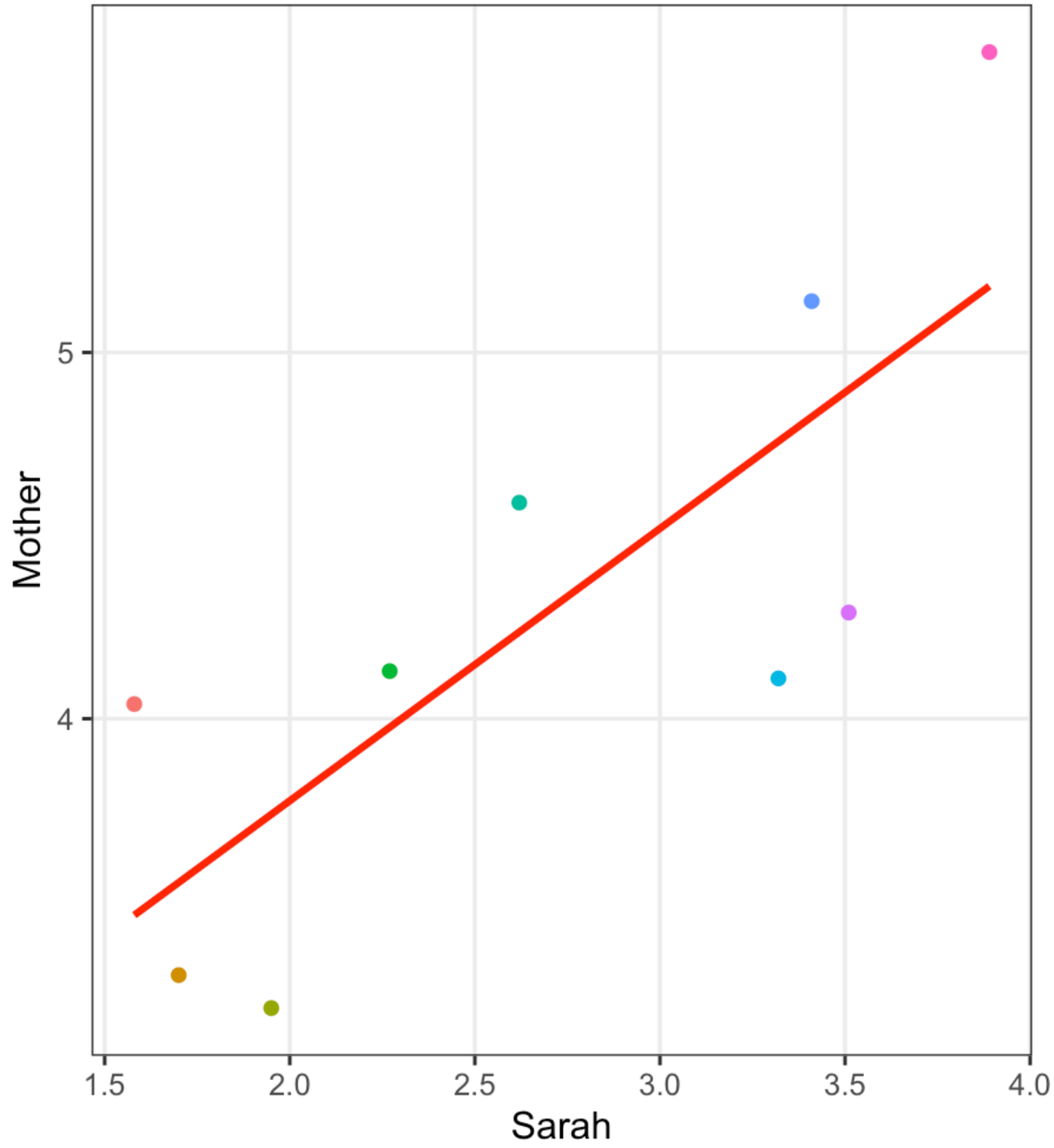
time	Sarah	Mother
06/1963	1.95	3.21
12/1963	1.58	4.04
02/1964	1.7	3.3
09/1964	2.27	4.13
12/1964	2.62	4.59
03/1965	3.32	4.11
06/1965	3.51	4.29
09/1965	3.89	5.82
12/1965	3.41	5.14

From covariance to correlation

- Covariance coefficient is too dependent on scale/units of measurement
  - → hard to interpret/generalize
- Solution: we **standardize (= z-score transform)** the covariance by dividing it by the product of the s of the two variables
- *Pearson's correlation coefficient*:
- $$r = \frac{cov(x, y)}{s_x s_y}$$

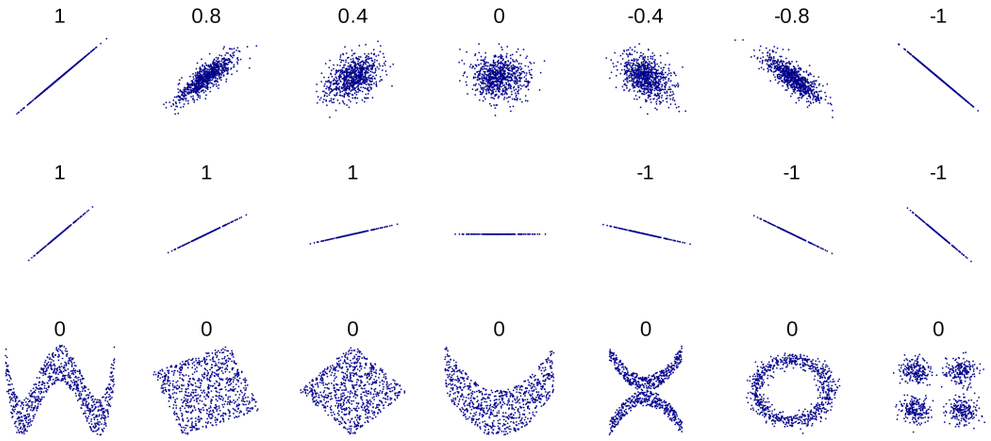
Pearson's Correlation Coefficient (1)

- $s_{Sarah} = 0.86$
- $s_{Mother} = 0.82$
- $r = \frac{0.554}{0.86 \times 0.82} =$
- $= \frac{0.554}{0.705} =$
- $= 0.78$



Pearson's Correlation Coefficient (2)

- Varies from -1 to +1
- r, ρ
- Higher value ≠ more positive slope
- BUT higher value = better fit (points are closer to the correlation line)

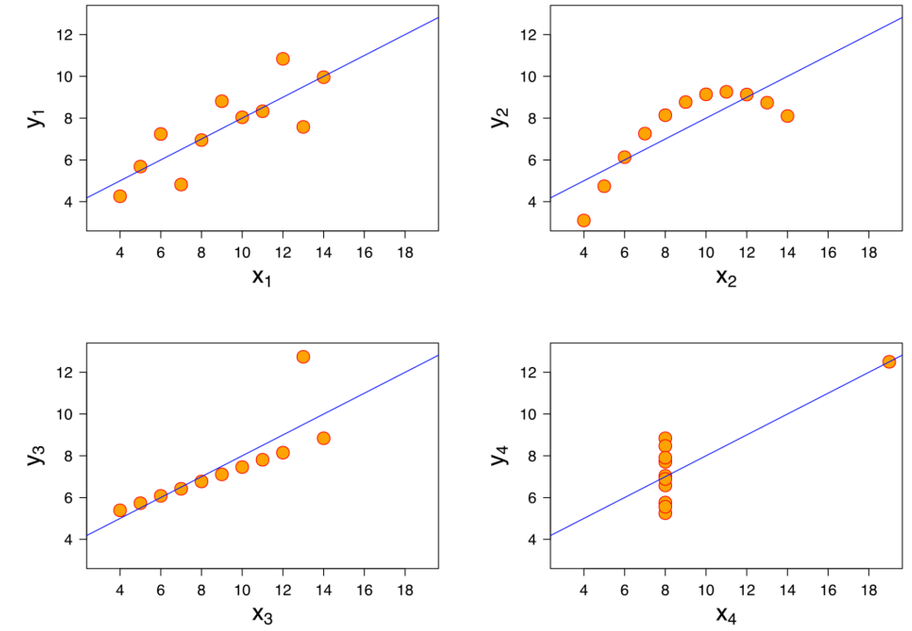


Value	Meaning
±1	Perfect relationship
±0.7	Strong relationship
±0.5	Moderate relationship
±0.3	Weak relationship
0	No (linear) relationship

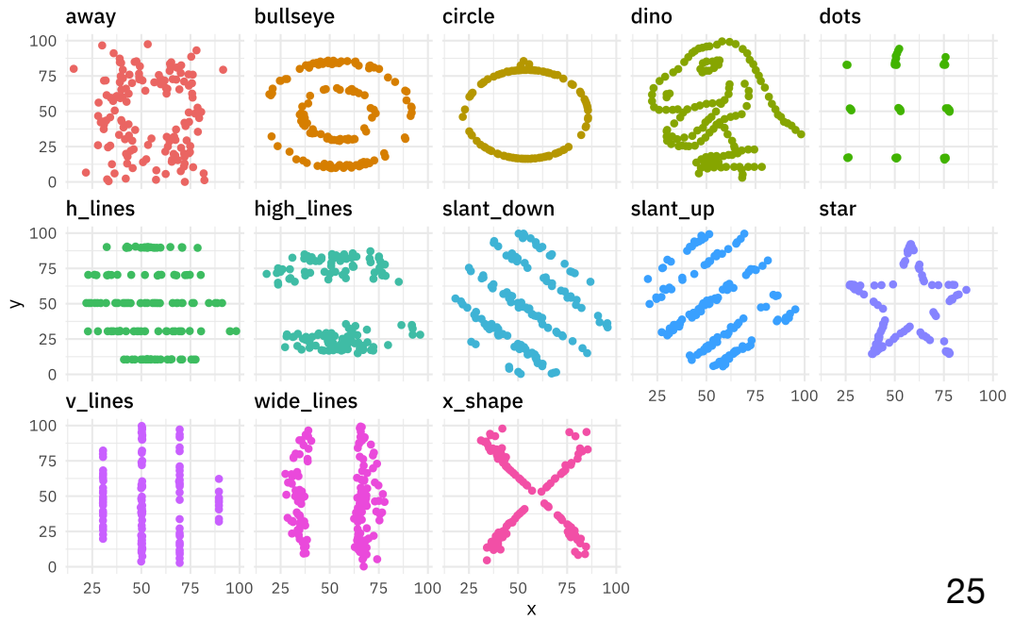
Bonus slide: Anscombe's Quartet

- Four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed
- Used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets
- Read more: https://www.wikiwand.com/en/Anscombe%27s_quartet

Anscombe's Quartet



The Datasaurus Dozen



Exercise… again

- Calculate the correlation coefficient for Sarah's and her mother's MLU

time	Sarah	Mother
06/1963	1.95	3.21
12/1963	1.58	4.04
02/1964	1.7	3.3
09/1964	2.27	4.13
12/1964	2.62	4.59
03/1965	3.32	4.11
06/1965	3.51	4.29
09/1965	3.89	5.82
12/1965	3.41	5.14

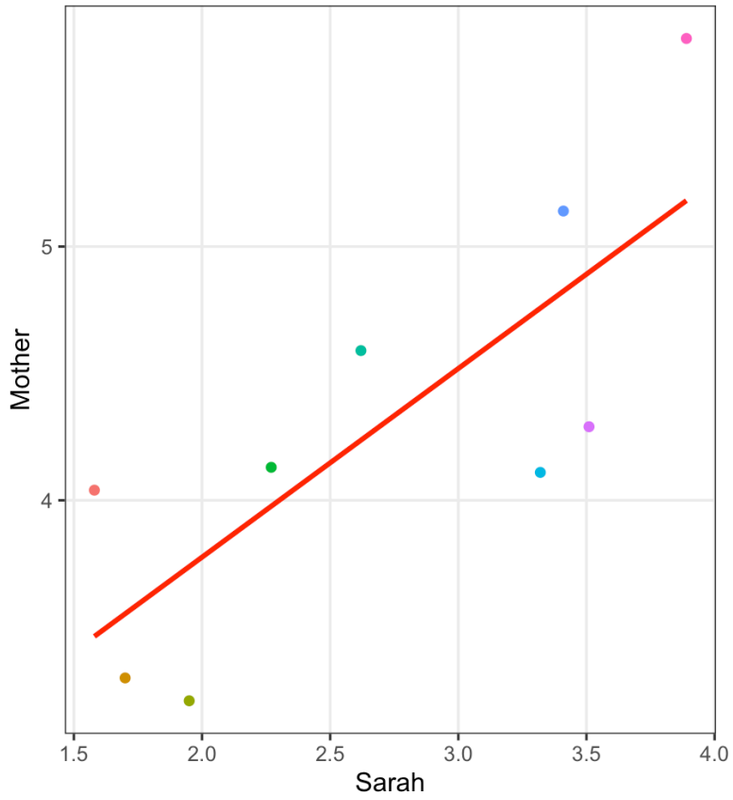
Correlation as measure of “effect size”

- = A standardized measure of the magnitude of the observed effect

- ie. What is the reciprocal effect of Sarah and her mom’s MLU on each other?

- $r_{(Sarah,\ Mother)} = 0.78$

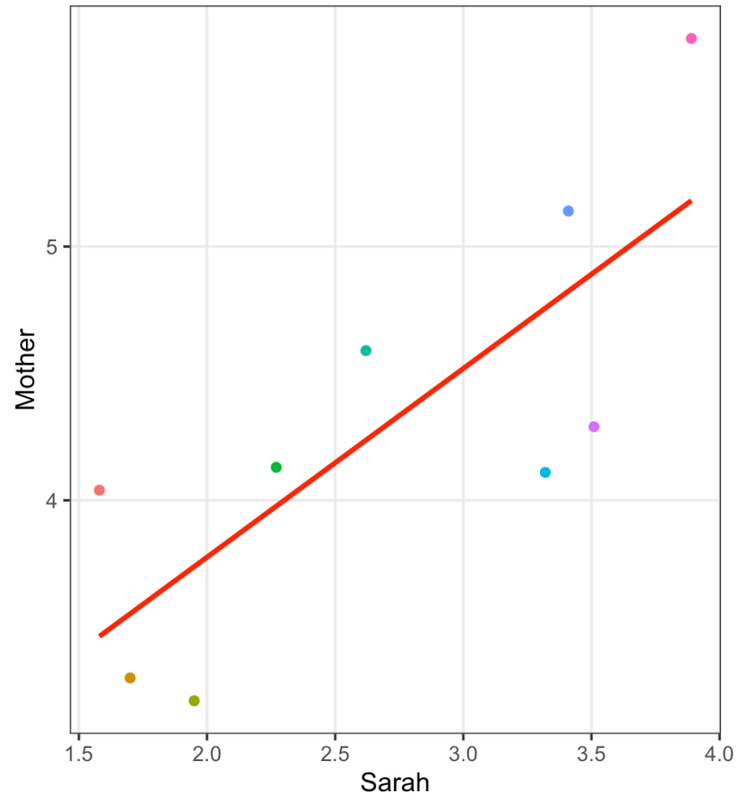
- Hard to interpret, not proportional



Effect size	*r*	Meaning
Small	0.1 < x < 0.3	Relationship explains 1-9% of total variance
Medium	0.3 < x < 0.5	Relationship explains 9%-25% of total variance
Large	x > 0.5	Relationship explains > 25% of total variance

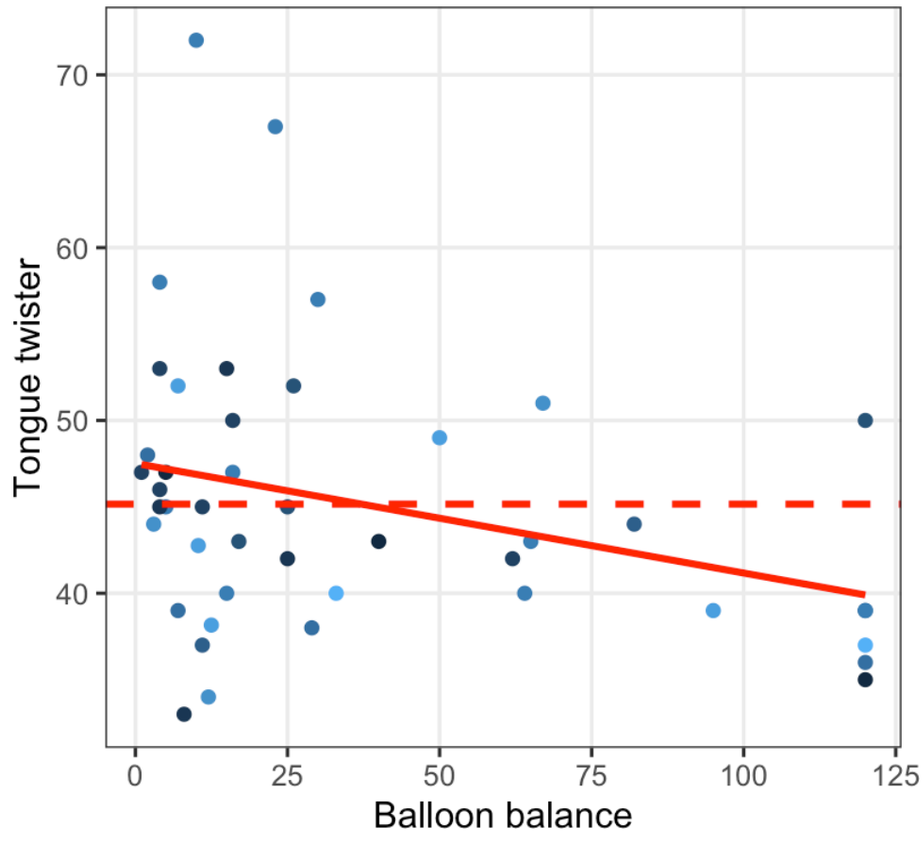
Coefficient of determination R^2

- = Measure of the amount of variability in one variable that is shared by another variable
  - r = 0.5 is not twice r = 0.25
  - but R^2 = 0.5 is twice R^2 = 0.25
- $R^2_{(Sarah, Mother)}$ = 0.78 * 0.78 = 0.61
- →61% of the total variance in our data is explained by the relationship between Sarah and her mom's MLUs
- Also a measure of goodness of fit, but more on this in Lecture 7 (Linear regression)



The *p*-value

- How can we determine whether the observed *r* is a true result or just noise?
- *p*-value = P(0,1) of getting an *r* value equal or higher than the observed one if the H_0 was true
- H_0: $\rho = 0$
- If $p < 0.05$, we consider *r* to be significant (high confidence in the results)
- More on *p*-values next week




```
> cor.test(df$balloon_balance, df$tongue_twist)

        Pearson's product-moment correlation

data:  df$balloon_balance and df$tongue_twist
t = -2.1617, df = 42, p-value = 0.03639
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.56063660 -0.02156788
sample estimates:
       cor
-0.3164222
```

Assumptions in correlation analysis

- ***Pearson's product-moment correlation coefficient***: parametric test, assumes normality, homoschedasticity, and continuous data

- If dealing with data that does not match the assumptions, you can:

    1. transform the data (log, square-root, etc.)

    2. use ***Spearman's rank correlation coefficient*** (ρ):

        - non-parametric, works with ordinal data

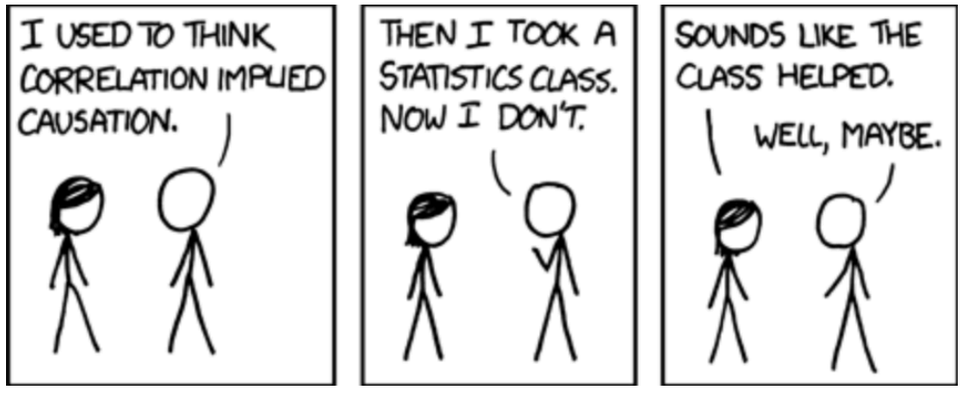
    3. use ***Kendall's rank correlation coefficient*** (τ):

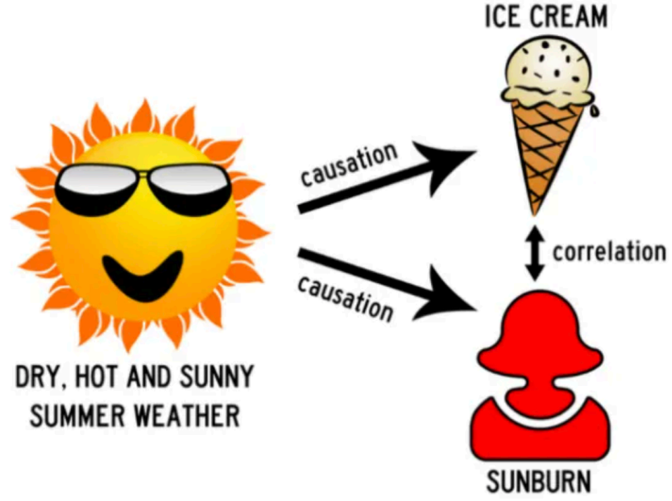
        - non-parametric

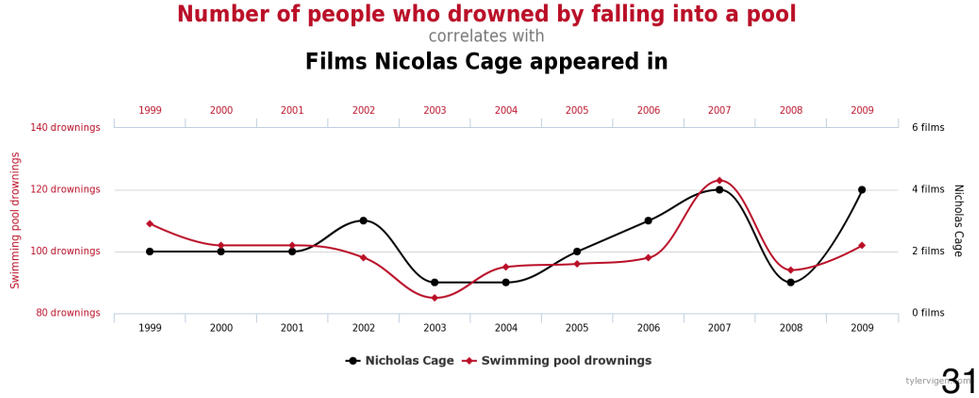
        - especially good for small samples

Potential issues with correlation analysis

- Direction of causality:
  - Correlation coefficients say nothing about which variable causes the other to change
  - Solution: the Randomized Control Trial
- The third/lurking variable problem:
  - Causality between two variables cannot be assumed because there may be other measured or unmeasured variables affecting the results.
  - Solution: causal structures/directed acyclic graphs
- Significant correlations that are not theoretically motivated/justified
  - Solution: always know your theory!







Thursday

- We will do correlation analysis on the data that you have collected for the PsychoPy workshop

- How do different variables affect our reading pace?

- Part of your Portfolio 2