BMR 617

Correlation

March 9th 2021

Q →Q case

- So far we have seen examples where both the explanatory variable and the response variable are categorical (C → C)
  - Vaccine example
  - Use contingency tables, look at relative risk, etc.
- And where the explanatory variable is categorical and the outcome is quantitative (C → Q)
  - Mouse data
  - Use boxplots, column scatter plots and/or bar charts
  - Measure mean, standard deviation, standard error of the mean, etc.
- Now we'll consider the case where both the explanatory variable and the response variable are quantitative (Q → Q)

Example

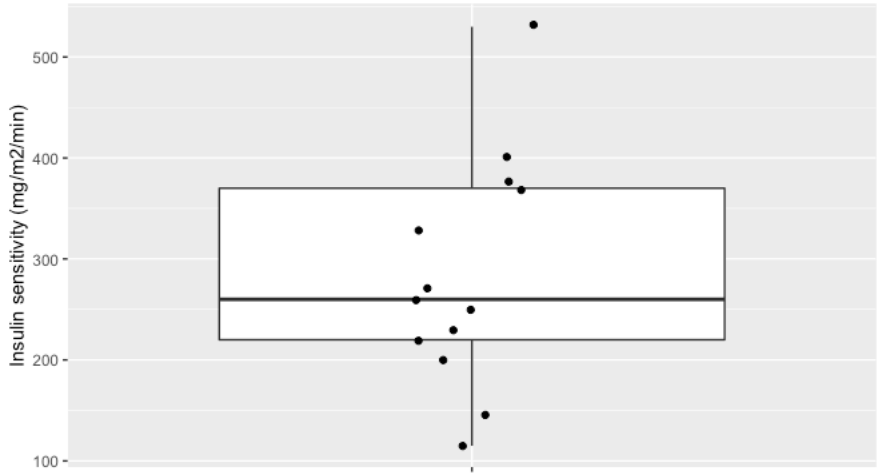
- Borkman et al. (New England Journal of Medicine, 1993) studied the relationship between insulin sensitivity and lipid composition of the cell membrane
- They measured insulin sensitivity of 13 healthy men by infusing insulin and measuring how much glucose they needed to infuse to maintain a constant blood glucose level
- The also took skeletal muscle biopsies and measured (among other things) the percentage of polyunsaturated fatty acids that had between 20 and 22 carbon atoms (%C20-22).

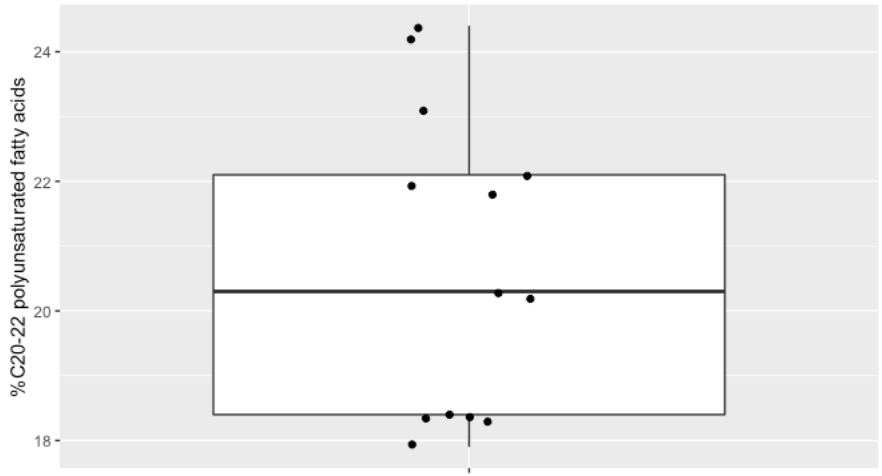
Data

Insulin sensitivity (mg/m^2/min)	%C20-22 polyunsaturated fatty acids
250	17.9
220	18.3
145	18.3
115	18.4
230	18.4
200	20.2
330	20.3
400	21.8
370	21.9
260	22.1
270	23.1
530	24.2
375	24.4

Variability

- Both these variables show a degree of variability:





Covariability or Correlation

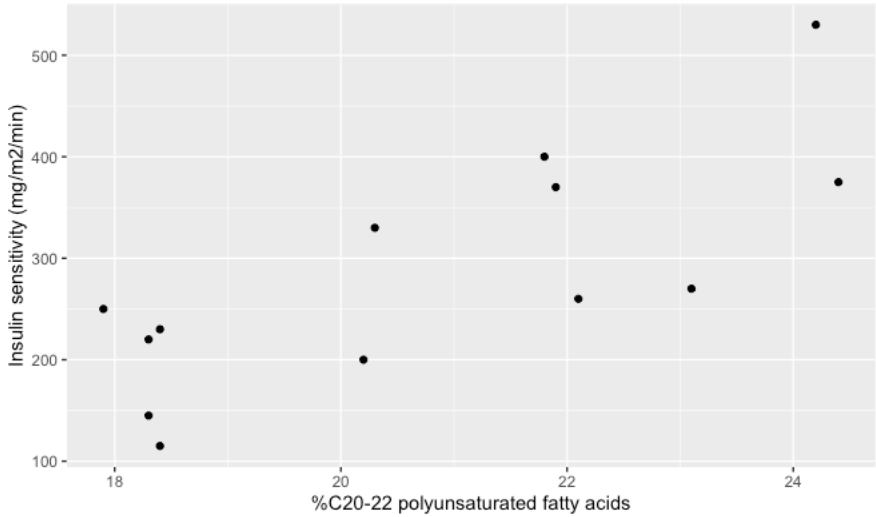
- If we plot both variables together in a scatterplot, we see that the variation is "shared" between the variables

```
library(tidyverse)

ins <-
read_csv("https://denvirlab.marshall.edu/BMR617-2021/data/InsulinSensitivityBorkman.csv")

ggplot(ins, aes(x=`%C20-22`, y=InsulinSensitivity)) + geom_point()
```

- Individuals who have more C20-22 polyunsaturated fatty acids tend to have higher insulin sensitivity
- We say there is a lot of covariability or correlation
- The variables "vary together"

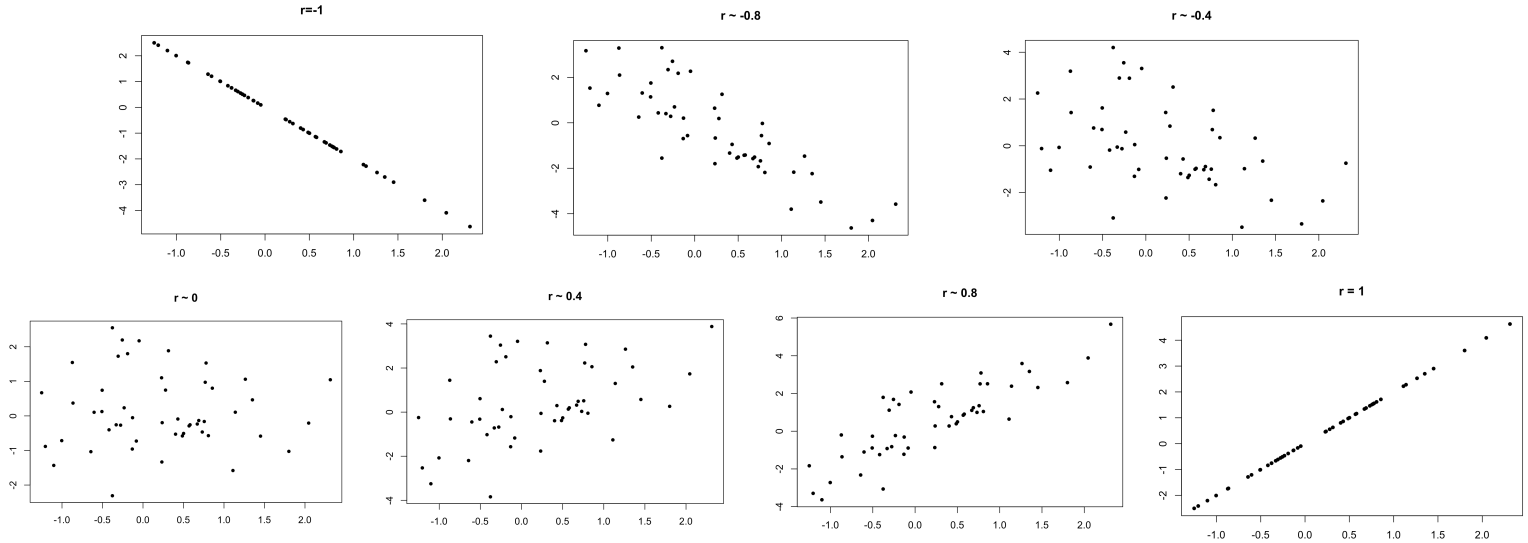


The correlation coefficient

- The amount of correlation can be quantified by a *correlation coefficient*
- The correlation coefficient between two sets of values $x_1 \ldots x_n$ and $y_1 \ldots y_n$ is computed as follows:
  - Calculate the *standardized values* of x and y:
  - $z_{x,i}=(x_i\text{-mean}(x))/\text{sd}(x)$; $z_{y,i}=(y_i\text{-mean}(y))/\text{sd}(y)$;
  - Compute the products of all the standardized values, add them up, and divide by n-1:
  - $r=(z_{x,1}z_{y,1}+z_{x,2}z_{y,2}+\ldots+z_{x,n}z_{y,n})/(n\text{-}1)$

Correlation

- The correlation coefficient between a set of pairs of quantitative values is a measure of the strength of a *linear* relationship between them
  - 0 means no linear relationship
  - 1 means a perfect positive linear relationship
  - -1 means a perfect negative linear relationship



Why the correlation coefficient works

- If a value is bigger than the mean, its standardized score is positive, otherwise its standardized score is negative
- The product of two standardized scores will be positive if both scores are positive, or both scores are negative i.e. if both scores are bigger than the mean, or both are less than the mean
- So if one variable tends to increase when the other tends to increase, the bulk of the products of standardized scores will be positive, and the correlation coefficient will be high
- On the other hand, if one variable tends to decrease when the other increases, the bulk of the products of standardized scores will be negative, and the correlation coefficient will be low
- If there is no relationship, the standardized scores will be randomly distributed, and their products will tend to cancel out

Computing the correlation coefficient in R

- In R, we can do the following:

```
cor(pull(ins, `%C20-22`), pull(ins, InsulinSensitivity))
```

- The correlation coefficient for these data is 0.77
- This indicates a strong positive correlation between the two variables

Interpreting the correlation coefficient

- It's usually easier to interpret the square of the correlation coefficient, which we write as R^2
- Here R=0.77, so $R^2 = 0.593$
- The interpretation of R^2 is that it is the proportion of variation that is shared between the two variables
- I.e. 59.3% of the variation in insulin sensitivity is associated with the variation in lipid content
  - The remaining 40.7% of the variation in insulin sensitivity is explained by other factors
- When dealing with correlation, this is symmetrical, so we can also say that 59.3% of the variation in lipid content is associated with the variation in insulin sensitivity

Correlation and Causation

- There are several possible explanations of the correlation we observe:

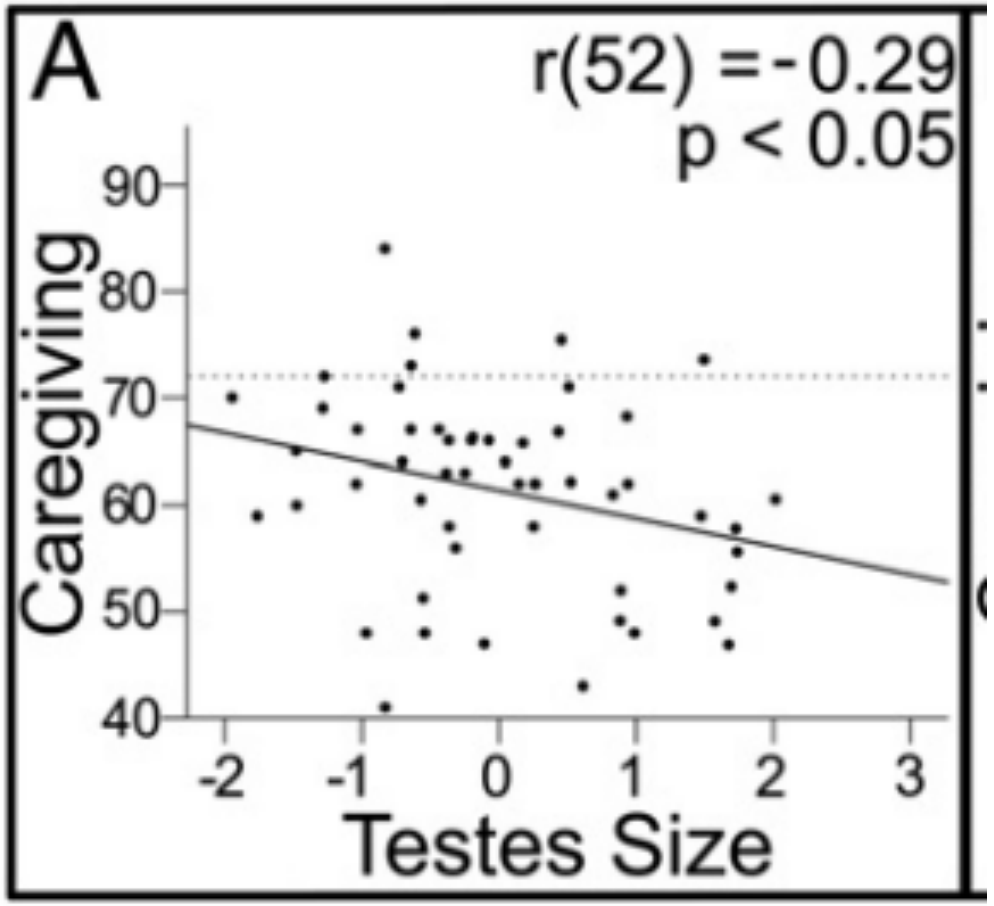
1. The lipid content of the membranes determines insulin sensitivity
2. The insulin sensitivity of the membranes affects its lipid content
3. Both lipid content and insulin sensitivity are under the control of another factor
4. Lipid content, insulin sensitivity, and several other factors are all part of a complex molecular network.
5. The two variables don't really correlate, and the observation in this sample was just a coincidence

Which explanation is correct?

- We'll learn how to quantify (to some degree) the likelihood of the last possibility later
- The only way to decide among the first four options is to do further experiments, where the variables are directly manipulated to see which variable affects which other variable(s)

Caution 1: beware of large samples

- Mascaro et al. (PNAS, 2013) published a study in which they measured a caregiving score and testes size of 52 fathers of children aged 1-2 years
- They computed a negative correlation between the two variables
- We'll discuss the meaning of p<0.05 later in the course
- Note here the correlation coefficient R=-0.29
- What is R^2? What is the interpretation?



Caution 2: beware of confounding variables

- In our mouse data, we could look at correlation between the metabolic variables.
- For example, to look at the correlation between Triglycerides and Glucose, we would do:

```
met <- read_csv("https://denvirlab.marshall.edu/BMR617-2021/data/TH-B6-metabolic.csv")
met <- separate(met, MouseID, sep="-", into=c("Strain", "Diet", "ID"))


cor(pull(met, TG), pull(met, Glucose))
```

- This gives an apparently very strong correlation of 0.878

A closer look at the Triglycerides-Glucose data

- Plot the data:

```
ggplot(met, aes(x=TG, y=Glucose)) +
    geom_point(aes(color=Diet, shape=Strain))
```

- The correlation is "driven" by a few data points which have much higher Triglycerides and Glucose
- Those data points are all TH mice on either the LF or HF diets
- Within any group, there is little to no evidence of correlation
- So it is likely that Strain and Diet affect both variables, but otherwise the two are unrelated