Computational
Social Science

Introduction to Monday's Analytics Oasis

University of Amsterdam

Welcome Analytics Oasis

oasis

noun

UK /əʊˈeɪ.sɪs/ US /oʊˈeɪ.sɪs/

plural **oases**

Add to word list

Delta Images/Image Source/GettyImages

[C]

a place in a desert where there is water and therefore plants and trees and sometimes a village or town

[S]

a calm, pleasant place in the middle of somewhere busy and unpleasant:

- *Her office was an oasis **of** peace and sanity amid the surrounding chaos.*

Welcome to Analytics Oasis

- We are especially inspired by the second definition, and want that Analytics Oasis provides a space that helps you to bring clarity to the topic statistics and the data that you work with.

Some students do the branding

Computational Social Science

Analytics OASIS

UvA

Learning Team 3, group **Black Mambas** (Sasha, Daniel, Merijn, Wentao, Viktor)

Lecturer: Dr. Steve Pickering

Topics for the coming weeks

- Hypothesis development and multiple Regression
- T-test and chi-squared test

Program for the next half hour

- Statistics in Semester 1
- Statistics in Semester 2 (including in Analytics Oasis)
- Resources for the Individual and Group Challenges that you will be working on in Semester 2
- The first Group Challenge

Semester 1

Statistics sessions: during the Deprogramming camp

Lectures:

- Introduction to Survey Design - What is a survey and when is it used to test hypotheses?
- Best Practices on Designing a Questionnaire and Selecting Scales
- Survey Design & Sampling Strategy
- Regressions and Missing Values
- Mostly Harmless Statistics: On Inference & Uncertainty

Workshops:

- Workshop: Testing Hypotheses I + II
- Workshop: 'Six Honest Hypotheses'
- Workshop: Statistics Revision.

Assignments:

- Assignment: Analyzing survey data in your Group Projects (graphs, tables, regression, even t-tests)
- Assignment: Research proposal

Semester 1 Statistics

- This is where we are coming from
- You can always return to the slides, literature, and Jupyter Notebooks used in Semester 1
- They will help you to prepare for what is to come…
- Make sure that the foundation is there

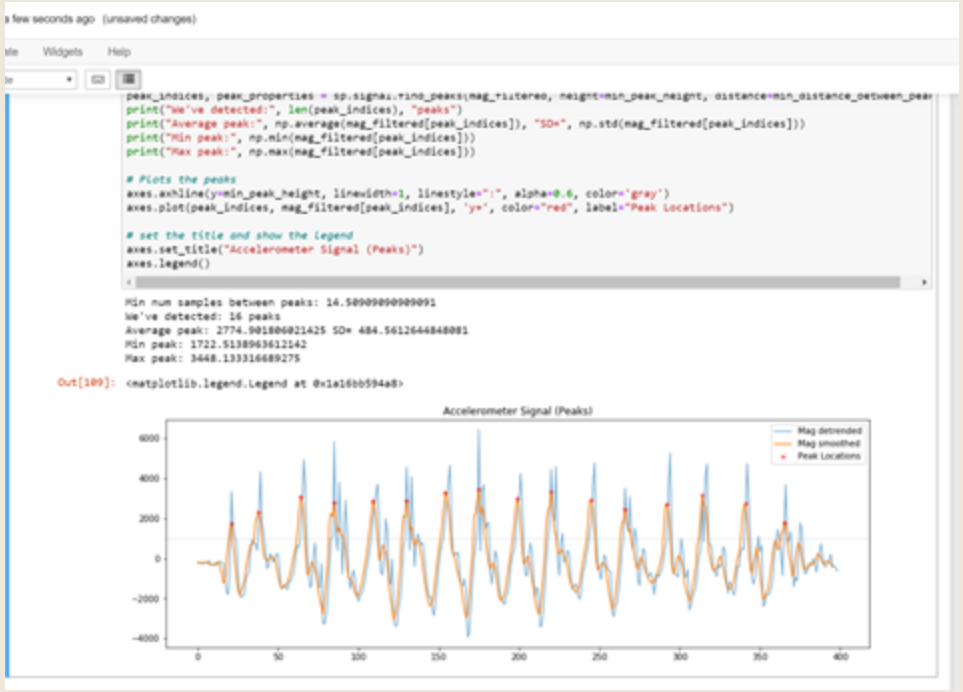
Semester 2: A portfolio that includes two Individual Challenges

- “In this individual assignment, you will complete a portfolio of Python notebooks that cover different statistical subjects. You will then be invited to an explainer session with your Core Lecturer during which you explain and justify your solutions to the notebook

Semester 2: A portfolio that includes two Individual Challenges

Individual Challenges

- 1. Hypothesis development and multiple Regression
- 2. T-test and chi-squared test



Key dates (see Individual Assignment RE)

Friday, May 9, 2025 (17:00)	**Final portfolio submission**
Monday, May 19, 2025 (11 – 13:00)	**Explainer session practice**
Week 15 and Week 16 Scheduled with your CL	**Explainer session**

Progress checks:

- Hypothesis development and multiple regression (Progress check **28 March 2025**, 17:00)
- T-test and chi-squared test (Progress check **25 April 2025**, 17:00)

In the coming week: Group Challenges

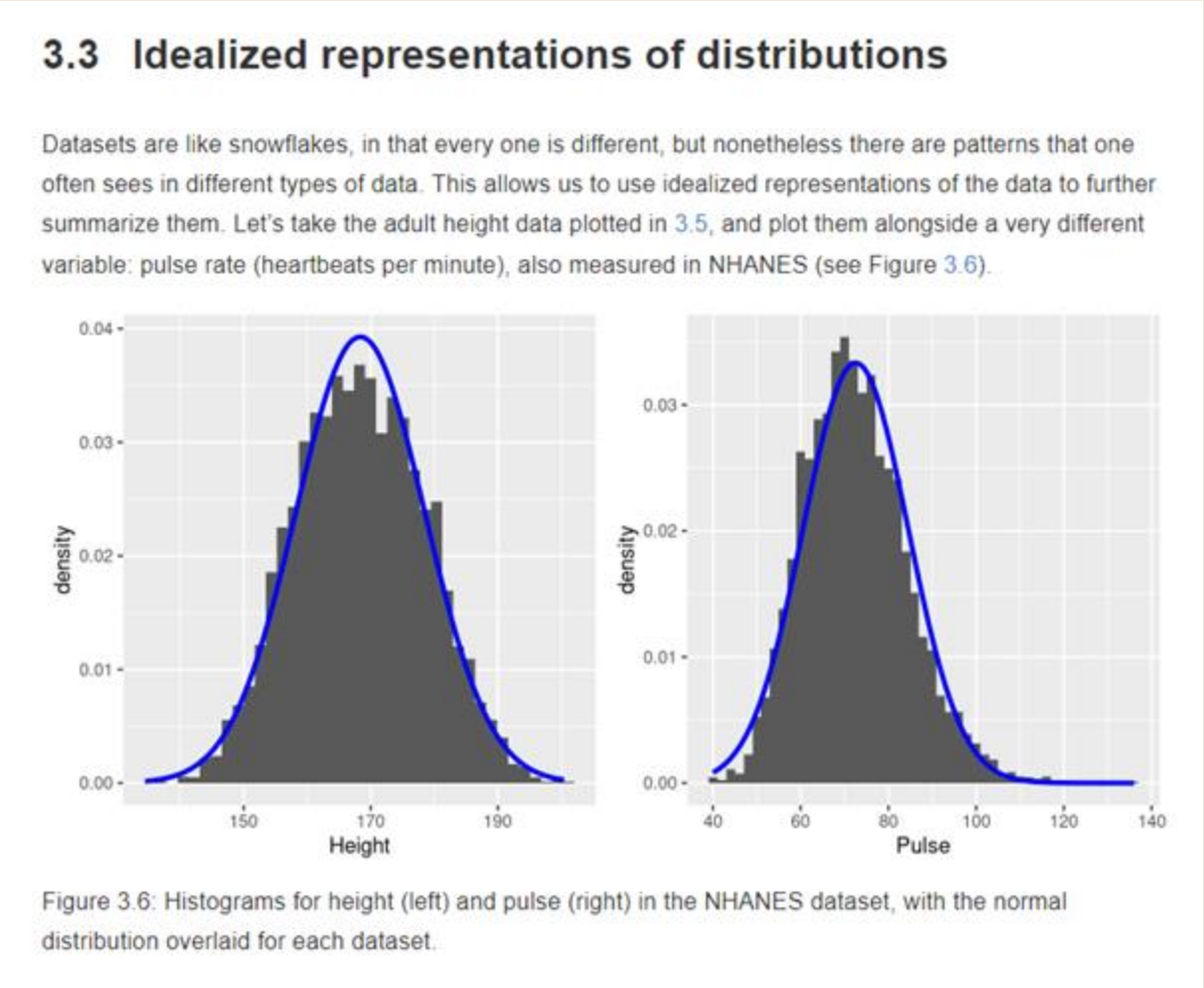
Group Challenges:

- Day 1 Group Challenge (Rens)
- Day 2 No Group Challenge
- Day 3 Group Challenge (Steve)
- Day 4 Group Challenge (Szilvi)
- Day 5 Group Challenge (Gizem)

Reward for daily Group Challenges: From the CSSci design studio

Resources for the Individual and Group challenges

- 1) Statistical Thinking for the 21st Century (for concepts)
- See also this separate book (for associated Python code).



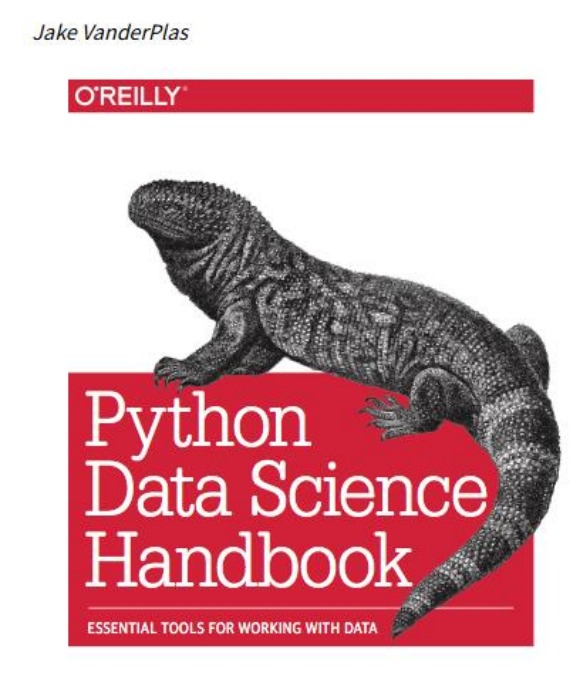
3.3 Idealized representations of distributions

Datasets are like snowflakes, in that every one is different, but nonetheless there are patterns that one often sees in different types of data. This allows us to use idealized representations of the data to further summarize them. Let's take the adult height data plotted in 3.5, and plot them alongside a very different variable: pulse rate (heartbeats per minute), also measured in NHANES (see Figure 3.6).

Figure 3.6: Histograms for height (left) and pulse (right) in the NHANES dataset, with the normal distribution overlaid for each dataset.

Resources for the Individual and Group challenges

- 2) Python Data Science Handbook



3. Data Manipulation with Pandas

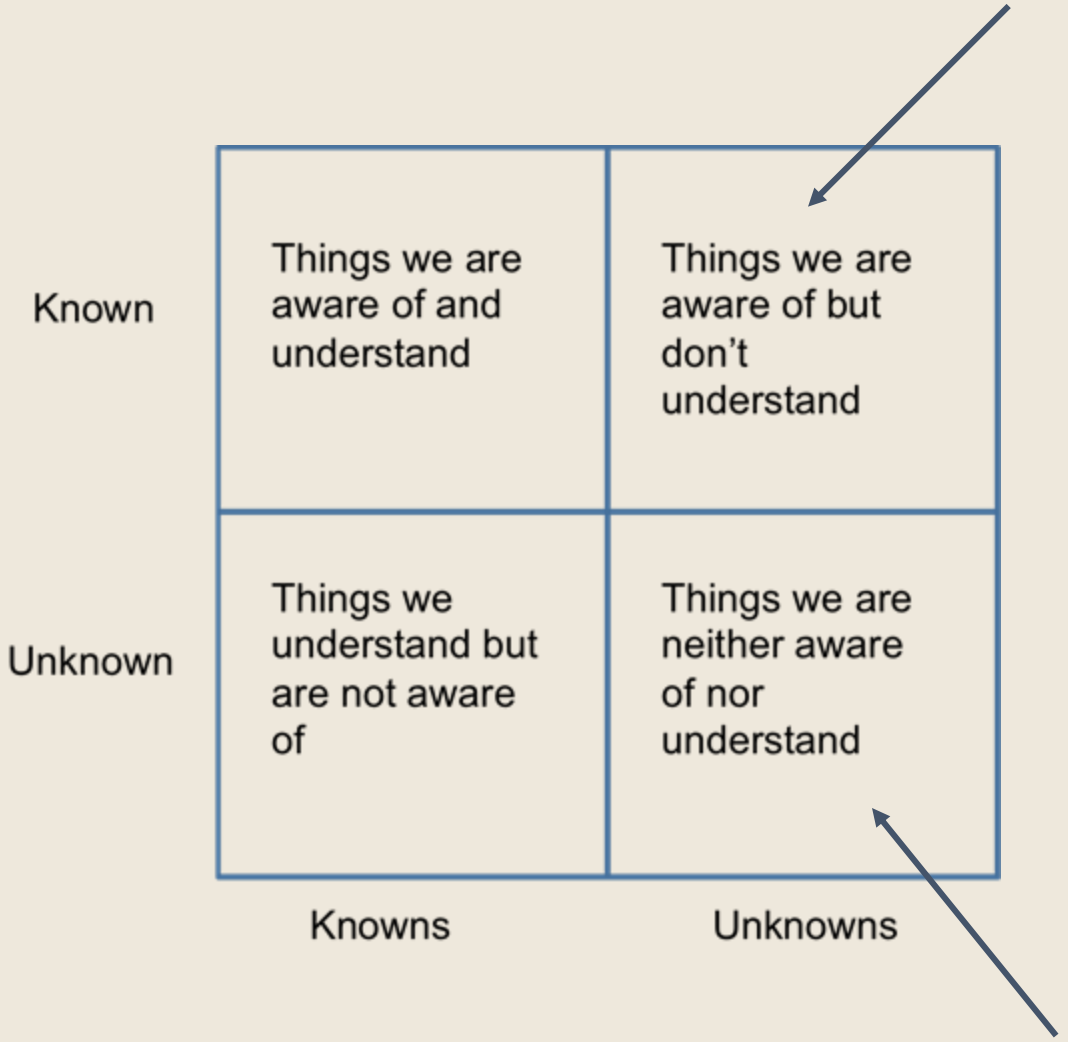
- Introducing Pandas Objects
- Data Indexing and Selection
- Operating on Data in Pandas
- Handling Missing Data
- Hierarchical Indexing
- Combining Datasets: Concat and Append
- Combining Datasets: Merge and Join
- Aggregation and Grouping
- Pivot Tables
- Vectorized String Operations
- Working with Time Series
- High-Performance Pandas: eval() and query()
- Further Resources

Resources for the Individual and Group challenges

- 3) Statistics by Jim
- Jim’s explanations often beat your average textbook explanation
- Jim’s website is an Analytics Oasis
- Jim deserves more recognition

Resources for the Individual and Group challenges

- All three resources can, of course, be used alongside Googling (and such)
- But beware, Googling → **Known - Unknowns**
- Textbooks and other resources → **Unknown - Unknowns**



Workshops, lectures, and the Analytics Oasis week

Besides yourself study, you will do the following support to prepare your portfolio:

- Statistics Review I + II
- Experiments I + II
- TA support
- Introduction lectures during the Analytics Oasis week

The Group Challenge

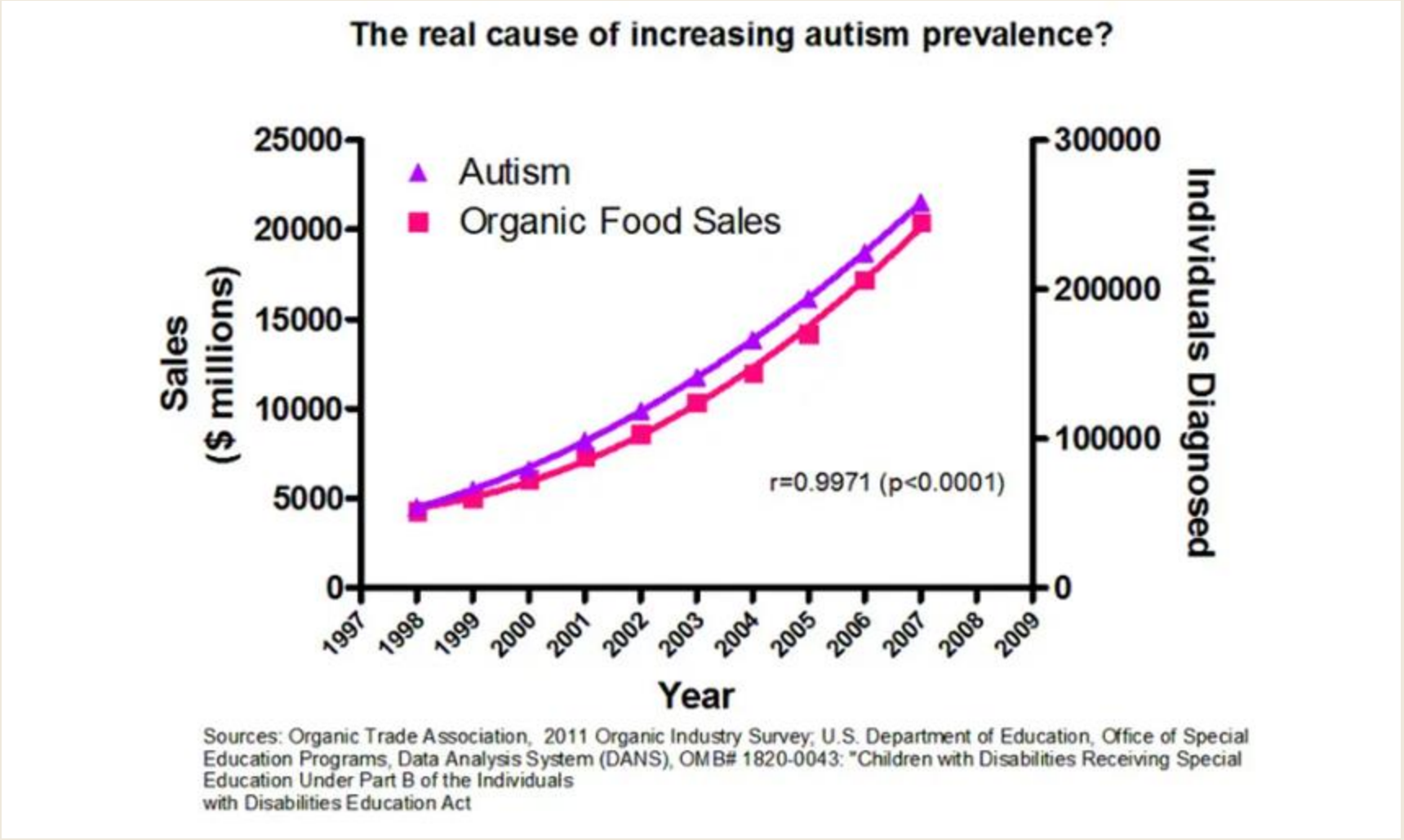
Monday’s Group Challenge

Produce and debunk bad science

You will be thrown into the deep end

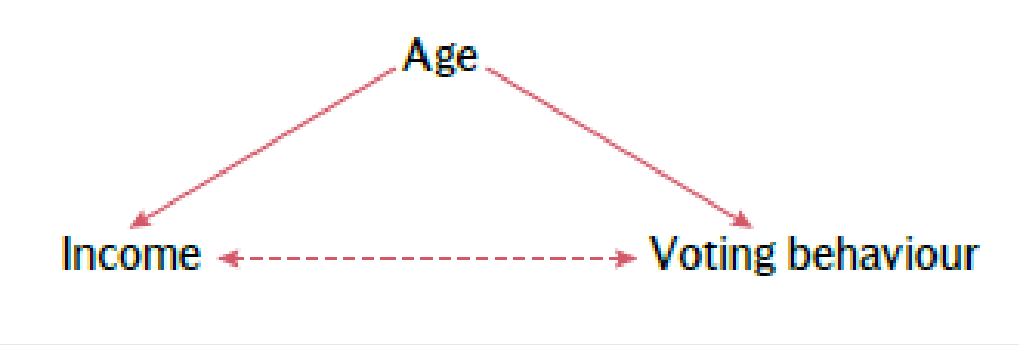
- Plotting, correlating, regressing hamburger restaurants with all kinds of nasty diseases or political preferences
- But I can also imagine that you sometimes will have the feeling “aren’t we going too fast?” - “Is it making sense what I’m doing now?”
- Today we are going to sensitize you to bad science by asking you to intentionally perform such practices

Is eating organic food causing autism



Spurious relationships

- “A spurious relationship exists when **there appears to be a relationship** between two variables, but the relationship is not real: it is being produced because each variable is itself related to a **third variable**” (Bryman, 2015).
- There appears to be a strong relation between income and voting behaviour (turnout)
- Could the relationship be an artefact of age? (the confounding variable)



Confounding variable?

- The relation between the number of firefighters and damage
- "... we might be tempted to conclude that the firefighters caused the property damage"

Number of Firefighters Called and Damage Caused (n=20)

	Few Firefighters	Many Firefighters	Total
Low Damage	7 (70%)	3 (30%)	10
High Damage	3 (30%)	7 (70%)	10
Total	10 (100%)	10 (100%)	20

Confounding variable?

Two partial tables showing the **size of the fire**

Small Fires (n=10)

	Few Firefighters	Many Firefighters	Total
Low Damage	4 (80%)	5 (100%)	9
High Damage	1 (20%)	0 (0%)	1
Total	5 (100%)	5 (100%)	10

Large Fires (n=10)

	Few Firefighters	Many Firefighters	Total
Low Damage	0 (0%)	1 (20%)	1
High Damage	5 (100%)	4 (80%)	9
Total	5 (100%)	5 (100%)	10

Could there be an intervening variable?

- Eelke's Dutch population study: Income → health
- Money on the bank by itself is not leading to a healthier life (or does it?)
- Income → access to good quality health care (among other things) → health
- Access to good quality health care = the **intervening variable**.
- (Side note: also called a **mediating variable)**.

Could there be a third moderating variable?

- Does the relationship between two variables hold for men but not for women?
- Or the other way around.
- If so, then the relationship is said to be “**moderated by gender**.”

Example: age and other sources of exercise

- Is there a relationship between **age** and **other sources of regular exercise** than going to the gym?
- Yes!

Table 15.6

Contingency table showing the relationship between age and whether or not gym visitors have other sources of regular exercise (%)

Other source of exercise	Age		
	30 and under	31–40	41 and over
Other source	64	43	58
No other source	36	57	42
n	42	23	24

Now we break it down by gender

Table 15.7

Contingency table showing the relationship between age and whether or not gym visitors have other sources of regular exercise for males and females (%)

Other source of exercise	Gender					
	Male			Female		
	30 and under	31–40	41 and over	30 and under	31–40	41 and over
Other source	70	33	75	59	50	42
No other source	30	67	25	41	50	58
n	20	9	12	22	14	12

… so, the relation between age and other sources of exercise is **moderated**, namely by gender.

Contingency tables

- Since there are potentially **confounding**, **intervening**, and **moderating variables**, we are advised to use **multiple regression** rather than bivariate statistics
- **Contingency tables** are useful basic technique to identify such spurious relationships.
- Alternatives: structural equation modelling and, of course, multiple regression (about which you will learn this semester)

Pandas (pd.crosstab) is super handy to make contingency tables

Confounding, intervening, and moderating variables

Type of variable	*Description*	*Basic way to identify it*
Confounding	(X ← Z → Y): it explains how two variables are affected by Z, e.g. size of fire (Z) affects both number of firefighters (X) and damage (Y).	**Correlation matrix and theory** (see Notebook). Or **contingency table** (the fire example)
Intervening	(X → Z → Y): it explains how X and Y are *connected*, e.g.: wealth (X) leads to a longer life span (Y), through the connector access to quality health care (Z).	**Correlation matrix and theory** (see Notebook).
Moderating	(X → Z → Y): the moderator *changes* the relation between X and Y, e.g. the effect of age (X) on other sources of exercise (Y), is varying with gender.	**Contingency table** (the exercise example). Or **correlation matrix and theory,** but with **splitted data** (e.g. one dataset for men and one for women).

The Group Challenge: produce and debunk bad science

- Intentionally **produce** bad science, e.g. by only focusing on two variables (rather than three) or by doing other bad science things (see the Notebook for examples)
- Then detect and **debunk** bad science by others

Monday’s Group Challenge

- Nicely share the tasks among team members
- You all have to open the dataset
- Some will run analyses
- Others will check the assumptions
- Others check which variables have the highest correlations

Materials are on Canvas

- There is a dataset that you can work with
- But you may also choose to work with another dataset (that you find yourself, such as on Kaggle)
- E.g.: the Traffic accidents dataset – or the Indian Marriage and Divorce Dataset
- I will be on the third floor in - or close to - the Common room

Important: by the end of the day

- **Post an online document in Year 1 Slack Channel**
- **Come to me** (on the third floor) to **pick up your Lego pieces**, while referring to your submission in the Slack Channel