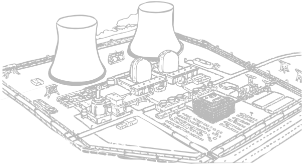


Modelling correlations with Python and SciPy

Eric Marsden

<eric.marsden@risk-engineering.org>

Context

- Analysis of **causal effects** is an important activity in risk analysis
  - Process safety engineer: *"To what extent does increased process temperature and pressure increase the level of corrosion of my equipment?"*
  - Medical researcher: *"What is the mortality impact of smoking 2 packets of cigarettes per day?"*
  - Safety regulator: *"Do more frequent site inspections lead to a lower accident rate?"*
  - Life insurer: *"What is the conditional probability when one spouse dies, that the other will die shortly afterwards?"*

- The simplest statistical technique for analyzing causal effects is **correlation analysis**

- Correlation analysis measures the extent to which two variables vary together, including the strength and direction of their relationship

Measuring linear correlation

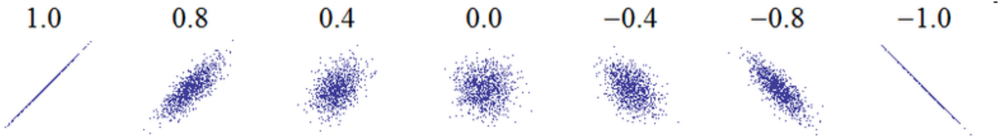
- **Linear correlation coefficient**: a measure of the strength and direction of a linear association between two random variables
  - also called the *Pearson product-moment correlation coefficient*
- $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$
  - $\mathbb{E}$ is the expectation operator
  - cov means covariance
  - μ_X is the expected value of random variable X
  - σ_X is the standard deviation of X
- Python: `scipy.stats.pearsonr(X, Y)`
- Excel / Google Docs spreadsheet: function `CORREL`

Measuring linear correlation

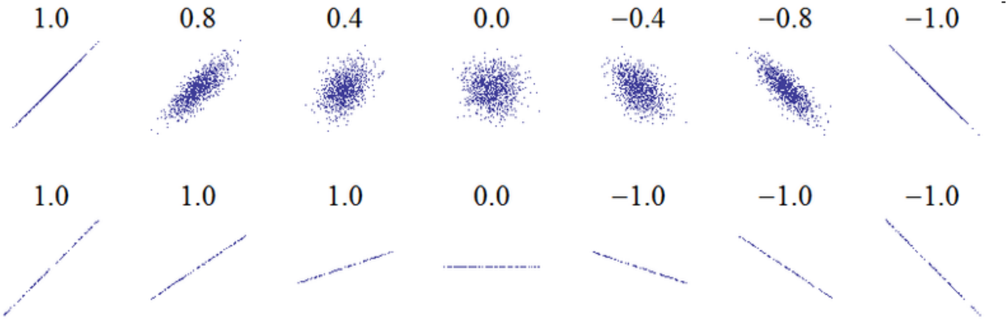
The linear correlation coefficient ρ quantifies the strengths and directions of movements in two random variables:

- **sign** of ρ determines the relative directions that the variables move in
- **value** determines strength of the relative movements (ranging from -1 to +1)
- ρ = 0.5: one variable moves in the same direction by half the amount that the other variable moves
- ρ = 0: variables are uncorrelated
  - *does not imply that they are independent!*

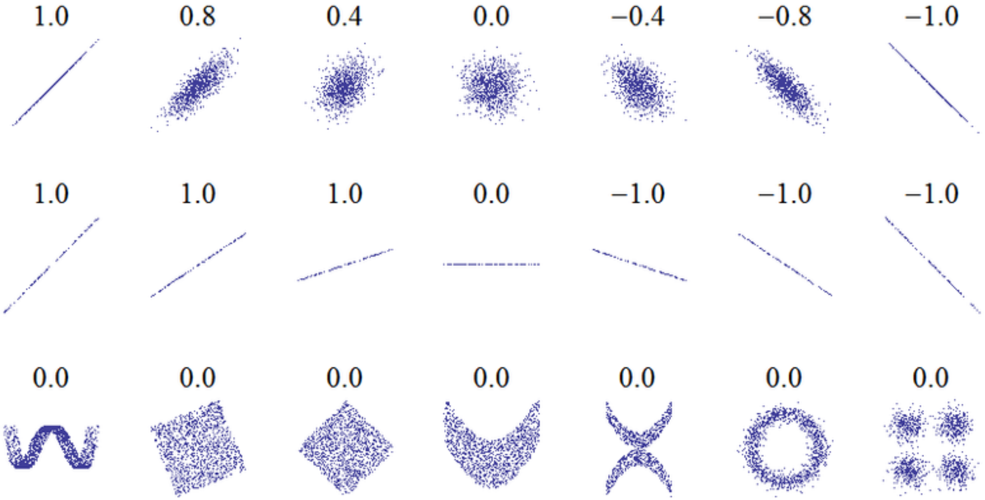
Examples of correlations



Examples of correlations



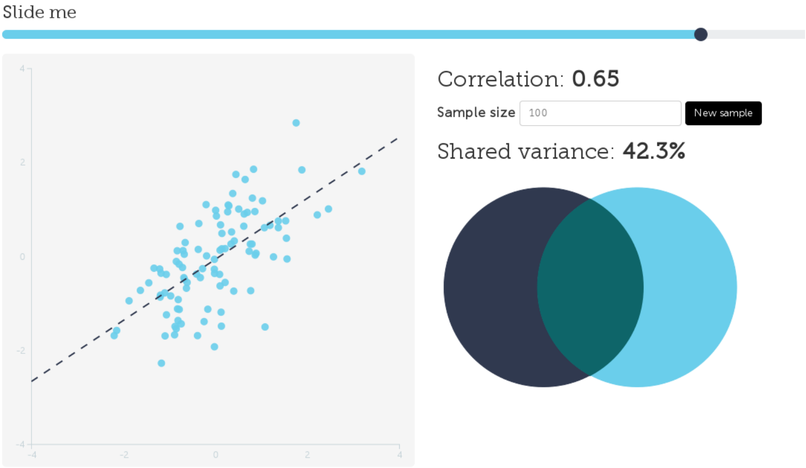
Examples of correlations



correlation ≠ dependency

Image source: Wikipedia

Online visualization: interpreting correlations



Correlation: **0.65**

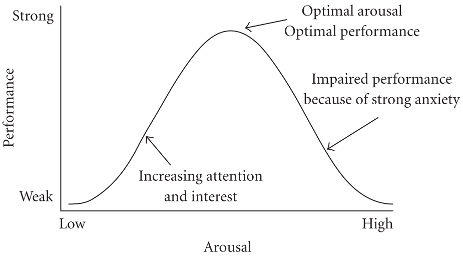
Sample size 100

Shared variance: **42.3%**

Try it out online: `rpsychologist.com/d3/correlation/`

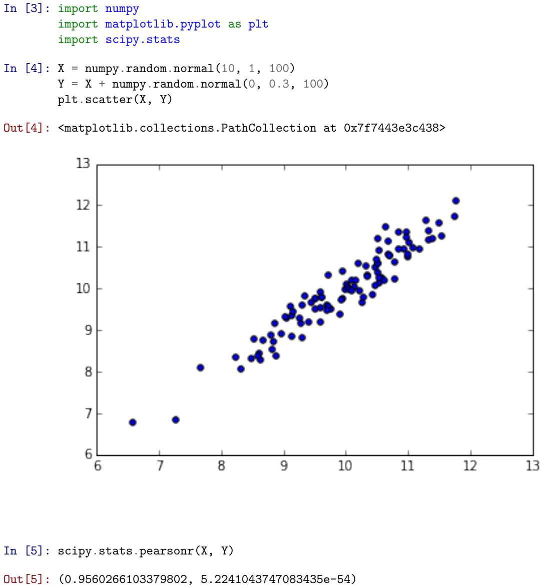
Not all relationships are linear!

- Example: Yerkes–Dodson law
  - empirical relationship between level of arousal/stress and level of performance
- Performance initially increases with stress/arousal
- Beyond a certain level of stress, performance decreases



Source: `wikipedia.org/wiki/Yerkes-Dodson_law`

Measuring correlation with NumPy



```
In [3]: import numpy
        import matplotlib.pyplot as plt
        import scipy.stats

In [4]: X = numpy.random.normal(10, 1, 100)
        Y = X + numpy.random.normal(0, 0.3, 100)
        plt.scatter(X, Y)

Out[4]: <matplotlib.collections.PathCollection at 0x7f7443e3c438>
```

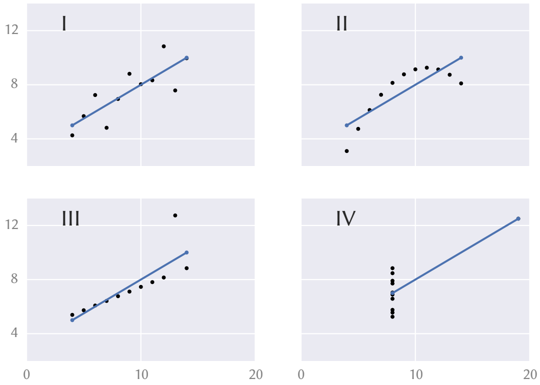
```
In [5]: scipy.stats.pearsonr(X, Y)

Out[5]: (0.9560266103379802, 5.2241043747083435e-54)
```

Exercise: show that when the error in Y *decreases, the correlation coefficient increases*

Exercise: produce data and a plot with a negative correlation coefficient

Anscombe's quartet

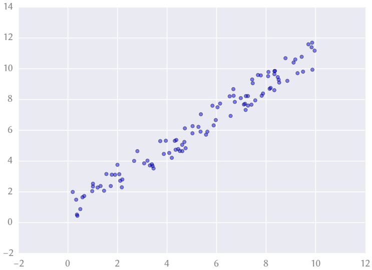


Each dataset has the same correlation coefficient!

Four datasets proposed by Francis Anscombe to illustrate the importance of **graphing data** rather than relying blindly on summary statistics

Plotting relationships between variables with matplotlib

- Scatterplot: use function `plt.scatter`
- Continuous plot or X-Y: function `plt.plot`



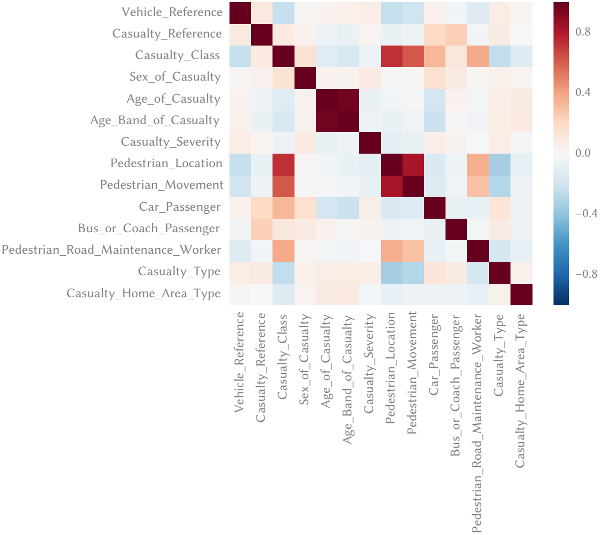
```
import matplotlib.pyplot as plt
import numpy

X = numpy.random.uniform(0, 10, 100)
Y = X + numpy.random.uniform(0, 2, 100)
plt.scatter(X, Y, alpha=0.5)
plt.show()
```

Correlation matrix

- A **correlation matrix** is used to investigate the dependence between multiple variables at the same time
  - output: a symmetric matrix where element m_{ij} is the correlation coefficient between variables i and j
  - note: diagonal elements are always 1
  - can be visualized graphically using a **correlogram**
  - allows you to see which variables in your data are informative
- In Python, can use:
  - `dataframe.corr()` method from the Pandas library
  - `numpy.corrcoef(data)` from the NumPy library
  - visualize using `imshow` from Matplotlib or `heatmap` from the Seaborn library

Correlation matrix: example



Analysis of the correlations between different variables affecting road casualties

```python
import pandas
import matplotlib.pyplot as plt
import seaborn as sns

data = pandas.read_csv("casualties.csv")
cm = data.corr()
sns.heatmap(cm, square=True)
plt.yticks(rotation=0)
plt.xticks(rotation=90)
```

Data source: UK Department for Transport, data.gov.uk/dataset/road-accidents-safety-data

Aside: polio caused by ice cream!

- ▷ Polio: an infectious disease causing paralysis, which primarily affects young children
- ▷ Largely eliminated today, but was once a worldwide concern
- ▷ Late 1940s: public health experts in USA noticed that the incidence of polio increased with the consumption of ice cream
- ▷ Some suspected that ice cream caused polio… sales plummeted
- ▷ Polio incidence increases in hot summer weather
- ▷ Correlation is not causation: there may be a hidden, underlying variable
  - *but it sure is a hint!* [Edward Tufte]

More info: *Freakonomics*, Steven Levitt and Stephen J. Dubner

Aside: fire fighters and fire damage

- Statistical fact: the larger the number of fire-fighters attending the scene, the worse the damage!
- More fire fighters are sent to larger fires
- Larger fires lead to more damage
- Lurking (underlying) variable = fire size
- An instance of “Simpson’s paradox”

Aside: low birth weight babies of tobacco smoking mothers

- Statistical fact: low birth-weight children born to smoking mothers have a *lower* infant mortality rate than the low birth weight children of non-smokers

- In a given population, low birth weight babies have a significantly *higher* mortality rate than others

- Babies of mothers who smoke are more likely to be of low birth weight than babies of non-smoking mothers

- Babies underweight because of smoking still have a lower mortality rate than children who have other, more severe, medical reasons why they are born underweight

- Lurking variable between smoking, birth weight and infant mortality

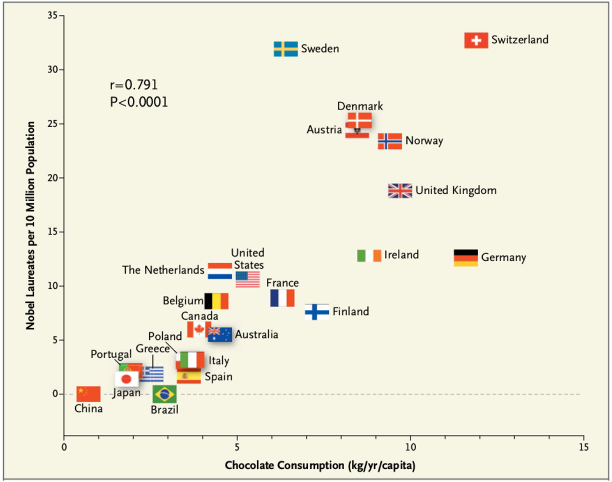
Source: Wilcox, A. (2001). *On the importance — and the unimportance — of birthweight*, International Journal of Epidemiology. 30:1233–1241

Aside: exposure to books leads to higher test scores

- In early 2004, the governor of the US state of Illinois R. Blagojevich announced a plan to mail one book a month to every child in in the state from the time they were born until they entered kindergarten. The plan would cost 26 million USD a year.
- Data underlying the plan: children in households where there are more books do better on tests in school
- Later studies showed that children from homes with many books did better even if they never read…
- Lurking variable: homes where parents buy books have an environment where learning is encouraged and rewarded

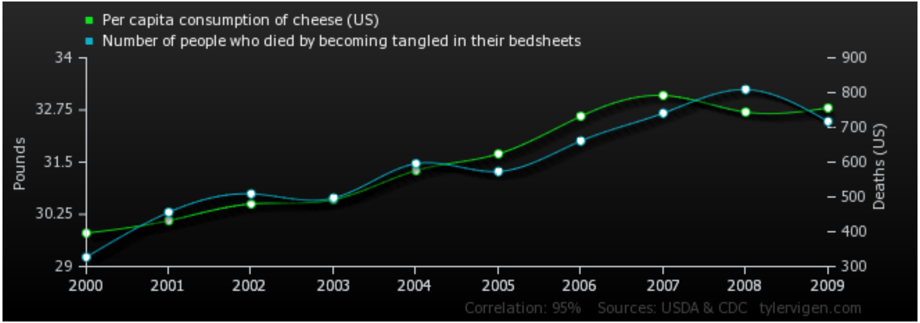
Source: `freakonomics.com/2008/12/10/the-blagojevich-upside/`

Aside: chocolate consumption produces Nobel prizes



Source: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012, DOI: 10.1056/NEJMon1211064

Aside: cheese causes death by bedsheet strangulation

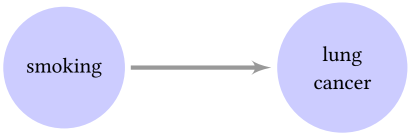


Note: real data!

Source: `tylervigen.com`, with many more surprising correlations

Beware assumptions of causality

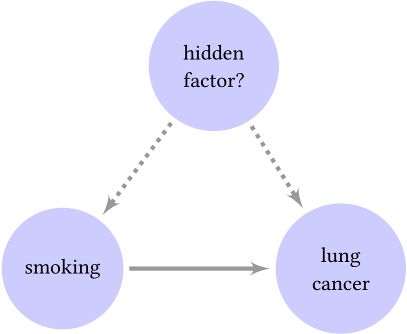
1964: the US Surgeon General issues a report claiming that cigarette smoking causes lung cancer, based mostly on correlation data from medical studies.



Beware assumptions of causality

1964: the US Surgeon General issues a report claiming that cigarette smoking causes lung cancer, based mostly on correlation data from medical studies.

However, correlation is not sufficient to demonstrate causality. There might be some hidden genetic factor that causes both lung cancer and desire for nicotine.



In logic, this is called the "post hoc ergo propter hoc" fallacy

Beware assumptions of causality

- To demonstrate the causality, you need a **randomized controlled experiment**
- Assume we have the power to force people to smoke or not smoke
  - and ignore moral issues for now!
- Take a large group of people and divide them into two groups
  - one group is obliged to smoke
  - other group not allowed to smoke (the "control" group)
- Observe whether smoker group develops more lung cancer than the control group
- We have eliminated any possible hidden factor causing both smoking and lung cancer
- More information: read about **design of experiments**

Constructing arguments of causality from observations

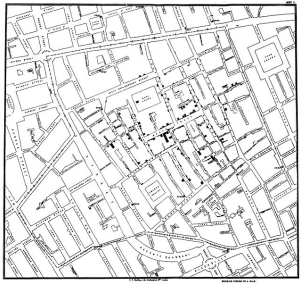
- Causality is an important — and complex — notion in risk analysis and many areas of science, with two main approaches used
- **Conservative approach** used mostly in the physical sciences requires
  - a **plausible physical model** for the phenomenon showing how A might lead to B
  - observations of correlation between A and B
- **Relaxed approach** used in the social sciences requires
  - a **randomized controlled experiment** in which the choice of receiving the treatment A is determined only by a random choice made by the experimenter
  - observations of correlation between A and B
- Alternative relaxed approach: a quasi-experimental "natural experiment"

Natural experiments and causal inference

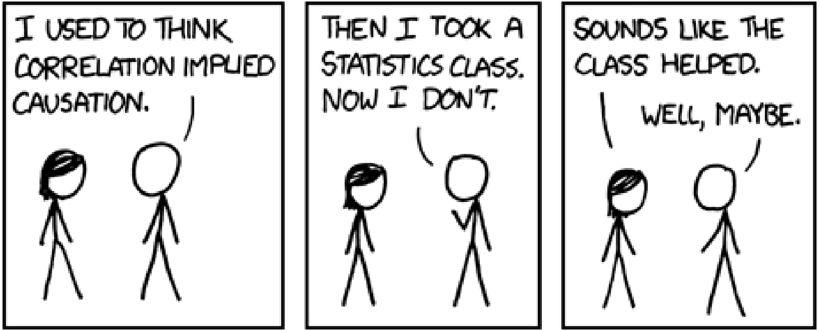
- **Natural experiment**: an empirical study in which allocation between experimental and control treatments are determined by factors outside the control of investigators but which resemble random assignment
- Example: in testing whether military service subsequently affected job evolution and earnings, economists examined difference between American males drafted for the Vietnam war and those not drafted
    - draft was assigned on the basis of date of birth, so “control” and “treatment” groups likely to be similar statistically
    - findings: earnings of veterans approx. 15% lower than those of non-veterans

Natural experiments and causal inference

- Example: cholera outbreak in London in 1854 led to 616 deaths
- Medical doctor J. Snow discovered a strong association between the use of the water from specific public water pumps and deaths and illnesses due to cholera
  - “bad” pumps supplied by a company that obtained water from the rivers Thames downstream of a raw sewage discharge
  - “good” pumps obtained water from the Thames upstream from the discharge point
- Cholera outbreak stopped when the “bad” pumps were shut down



Aside: correlation is not causation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Source: xkcd.com/552/ (CC BY-NC licence)

RISK ENGINEERING

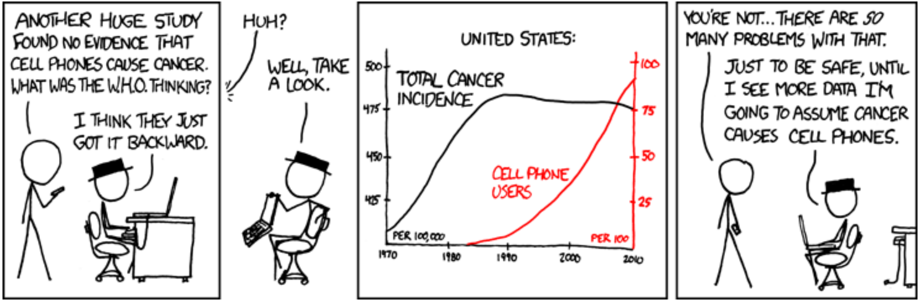
Directionality of effect problem

aggressive behaviour → watching violent films

aggressive behaviour ← watching violent films

Do aggressive children prefer violent TV programmes, or do violent programmes promote violent behaviour?

Directionality of effect problem


ANOTHER HUGE STUDY FOUND NO EVIDENCE THAT CELL PHONES CAUSE CANCER. WHAT WAS THE W.H.O. THINKING?
I THINK THEY JUST GOT IT BACKWARD.
HUH?
WELL, TAKE A LOOK.
UNITED STATES:
TOTAL CANCER INCIDENCE
CELL PHONE USERS
500
475
450
425
PER 100,000
100
75
50
25
PER 100
1970
1980
1990
2000
2010
YOU'RE NOT... THERE ARE SO MANY PROBLEMS WITH THAT.
JUST TO BE SAFE, UNTIL I SEE MORE DATA I'M GOING TO ASSUME CANCER CAUSES CELL PHONES.

Source: `xkcd.com/925/` (CC BY-NC licence)

Further reading

You may also be interested in:

- slides on **linear regression modelling using Python**, the simplest approach to modelling correlated data
- slides on **copula and multivariate dependencies for risk models**, a more sophisticated modelling approach that is appropriate when dependencies between your variables are not linear

Both are available from `risk-engineering.org` and from `slideshare.net/EricMarsden1`.

Image credits

- Eye (slide 21): Flood G. via `flic.kr/p/aNpvLT`, CC BY-NC-ND licence
- Map of cholera outbreaks (slide 23) by John Snow (1854) from Wikipedia Commons, public domain

For more free course materials on risk engineering, visit `risk-engineering.org`

For more information

- SciPy lecture notes: `scipy-lectures.github.io`
- Analysis of the “pay for performance” (correlation between a CEO’s pay and their job performance, as measured by the stock market) principle, `freakonometrics.hypotheses.org/15999`
- Python notebook on a more sophisticated Bayesian approach to estimating correlation using PyMC, `nbviewer.jupyter.org/github/psinger`

For more free course materials on risk engineering, visit `risk-engineering.org`

Feedback welcome!

Was some of the content unclear? Which parts of the lecture were most useful to you? Your comments to `feedback@risk-engineering.org` (email) or `@LearnRiskEng` (Twitter) will help us to improve these course materials. Thanks!

`@LearnRiskEng`

`fb.me/RiskEngineering`

`google.com/+RiskengineeringOrgCourseware`

For more free course materials on risk engineering, visit `risk-engineering.org`