Tutorial 3:

Introduction to Data Science

How to get Data?

Ernst C. Wit

September 29, 2024

Exercises

1. **What is an Astronomical Statistical Unit?** Youden (1994) collected data from various studies trying to determine the mean distance from the earth to the sun. This is called an *astronomical unit* (AU) and is fundamental in astronomy to describe distances. The distance from the Earth to the Sun is not constant. The Earth's orbit around the Sun is elliptical, meaning it varies slightly throughout the year. At its closest approach to the Sun (perihelion), Earth is about 91.4 million miles away, whereas at its farthest point (aphelion), Earth is about 94.5 million miles away from the Sun. The data is available as `AU.csv` on the course website.

	study	year	million miles
1	Newcombe	1895	93.28
2	Hinks	1901	92.83
3	Noteboom	1921	92.91
4	Spencer 1	1928	92.87
5	Spencer 2	1931	93.00
6	Witt	1933	92.91
7	Adams	1941	92.84
8	Brouwer	1950	92.98
9	Rabe	1950	92.91
10	Millstone Hill	1958	92.87
11	Jodrell Bank 1	1959	92.88
12	STL	1960	92.93
13	Jodrell Bank 2	1961	92.96
14	Caltech	1961	92.96
15	Soviets	1961	92.81

   a. What is the population in this experiment?

   b. What is the parameter of interest?

   c. How good is the sample from the population: evaluate *representativeness, bias and size*.

2. **Isaac Newton as a historian.** Newton's last book was, perhaps surprisingly, titled *The Chronology of Ancient Kingdoms Amended*, published posthumously in 1728. In it he tries to argue that earlier historians were wrong to argue that the average reign of ancient kings was between 35 to 40 years. Moreover, he wanted to make a serious point

that the average reign was a good way to calculate historical events, where precise dates were unavailable.

In the book he presented the following table, which is available as `kings.csv` on iCorsi.

	Kingdom	Number of kings	Years
1	Judah	18	390
2	Israel	15	259
3	Babylon	18	209
4	Persia	10	208
5	Syria	16	244
6	Egypt	11	277
7	Macedonia	8	138
8	England (1066–1714)	30	648
9	France (first 24)	24	458
10	France (second 24)	24	451
11	France (last 15)	15	315

(a) What is the population and parameter of interest?

(b) In Newton's sample, each epoch is a statistical unit with its own "average reign" data value. For each epoch calculate the average reign and store this in a vector `AR`.

(c) Use the vector `AR` to estimate the parameter of interest.

(d) Newton realised that the estimate of the parameter was probably a bit wrong. And in the book he writes that the reign of kings should be reckoned "at about eighteen or twenty years a-piece." (Stigler, 2002, p.395). Where did he get these numbers from?

i. Newton was a Master of the Mint and knew about the concept of *Remedy* as part of the Trial of the Pyx. This was the maximum allowed deviation of a particular coin from its standard weight. The Remedy was proportional to the modern concept of *standard deviation*:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

where x_i is the i-th observation and $\bar{x}$ is the average of the observations. Calculate the standard deviation of the 11 average reigns in the various epochs.

ii. Does the standard deviation explain Newton's observation?

iii. The standard deviation expresses the variability of 1 observation, whereas the concept of *standard error* describes the variability of a mean of n observations and it is defined as:

$$s.e. = \frac{s}{\sqrt{n}}.$$

Calculate the standard error for the 11 observed average reigns and compare it to Newton's expression of uncertainty.

3. **Consider your own data records.**

In the first tutorial, you created a dataset with at least two features for each statistical unit associated with the content of the following painting:

(a) What was your population and parameter(s) of interest?

(b) Mention positive and negative aspects of your sample.