Week 3: Statistical Modelling

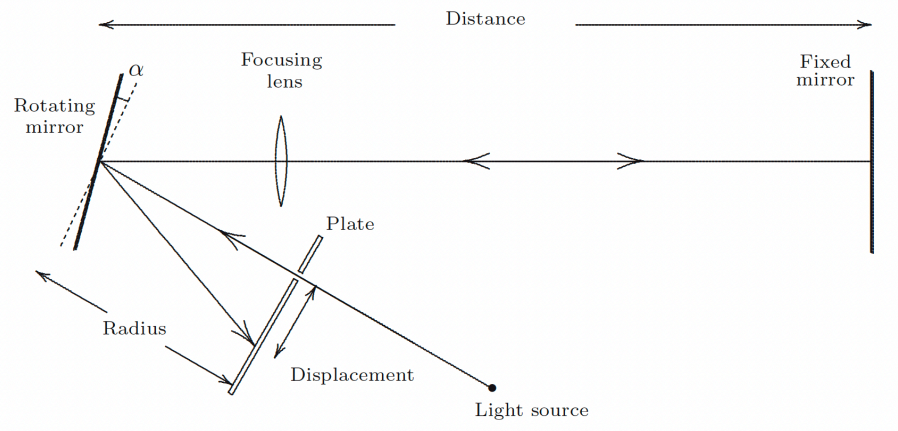
STA238: Probability, Statistics, and Data Analysis II

Luis Sierra Muntané

2026-01-22

Modelling example: the speed of light

- The data consists of measurements from 1879 to calculate the speed of light. i.e. measurements in units km/s - 299000 (km/s)
- See Section 1.6 and 17.1 of Dekking et al. (MIPS) for details.
- They are available as `morley` data set in `R`.



Modelling example: the speed of light

The table shows the last 3 digits of the measured speed of light for 100 measurements.

We can model the measurements as random variables $X_1, \ldots, X_{100}$.

The data in the table would correspond to realizations $x_1, \ldots, x_{100}$.

850	1000	960	830	880	880	890	910	890	870
740	980	940	790	880	910	810	920	840	870
900	930	960	810	880	850	810	890	780	810
1070	650	940	880	860	870	820	860	810	740
930	760	880	880	720	840	800	880	760	810
850	810	800	830	720	840	770	720	810	940
950	1000	850	800	620	850	760	840	790	950
980	1000	880	790	860	840	740	850	810	800
980	960	900	760	970	840	750	850	820	810
880	960	840	800	950	840	760	780	850	870

Speed of light measurements by Albert Michelson, 1879

Modelling example: the speed of light

850	1000	960	830	880	880	890	910	890	870
740	980	940	790	880	910	810	920	840	870
900	930	960	810	880	850	810	890	780	810
1070	650	940	880	860	870	820	860	810	740
930	760	880	880	720	840	800	880	760	810
850	810	800	830	720	840	770	720	810	940
950	1000	850	800	620	850	760	840	790	950
980	1000	880	790	860	840	740	850	810	800
980	960	900	760	970	840	750	850	820	810
880	960	840	800	950	840	760	780	850	870

Speed of light measurements by Albert Michelson, 1879

- Can this data be considered as realizations from a **random sample**?
  - Are the observations realizations of independent random variables?
    - Michelson designed the experiment so that individual measurements did not influence one another.
    - Each trial was conducted separately, with the apparatus reset between runs and measurements taken on different days (at sunrise).
  - It is therefore reasonable to assume **independence**.

Modelling example: the speed of light

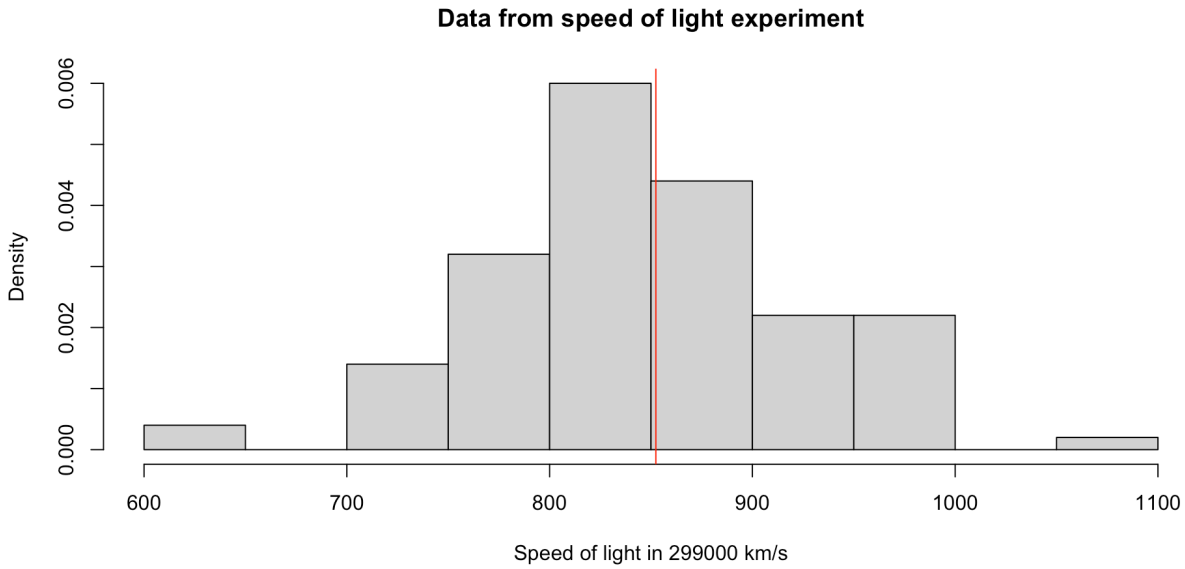
850	1000	960	830	880	880	890	910	890	870
740	980	940	790	880	910	810	920	840	870
900	930	960	810	880	850	810	890	780	810
1070	650	940	880	860	870	820	860	810	740
930	760	880	880	720	840	800	880	760	810
850	810	800	830	720	840	770	720	810	940
950	1000	850	800	620	850	760	840	790	950
980	1000	880	790	860	840	740	850	810	800
980	960	900	760	970	840	750	850	820	810
880	960	840	800	950	840	760	780	850	870

Speed of light measurements by Albert Michelson, 1879

- Can this data be considered as realizations from a **random sample**?
  - Are these observations realizations from the same data generating distribution?
    - The measurements were collected under the same experimental conditions.
    - Each measurement was intended to estimate the same underlying physical quantity (the speed of light).
  - It is therefore reasonable to assume **identically distributed**.

EDA on speed of light data

```r
mean <- mean(data)
hist(data, freq = FALSE,
     main = "Data from speed of light experiment",
     xlab="Speed of light in 299000 km/s")
abline(v = mean, col = "red")
```



What would be a suitable statistical model?

Modelling example: the speed of light

- Assume X_i = speed of light + measurement error
- Assume measurements are centered around the true speed of light μ_c: $\mathbb{E}[X_i] = \mu_c$
- Assume these measurement error are i.i.d. and normally distributed with mean 0 (no systematic error) and finite variance σ_2.
  - Assuming a symmetric error distribution
  - I.e. equally likely to underestimate or overestimate: $\mathbb{P}(X_i \leq \mu_c) = \mathbb{P}(X_i \geq \mu_c)$
- Thus the candidate statistical model is a normal distribution

$$X_1, \ldots, X_{100} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_c, \sigma^2)$$

Modelling example: the speed of light

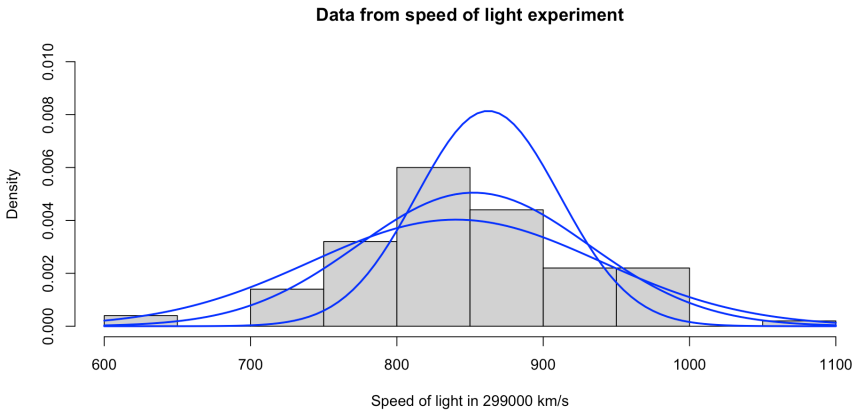
We choose the statistical model

$$X_1, \ldots, X_{100} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_c, \sigma^2)$$

- There are 2 model parameters: μ_c and σ^2
  - The true speed of light $\mu_c \geq 0$ (our main parameter of interest)
  - The variance $\sigma^2 > 0$ of measurement error.

The parameter space is then $[0, \infty) \times (0, \infty)$.

Defining the statistical model



Chosen statistical model:

$$X_1, \ldots, X_{100} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_c, \sigma^2)$$

for $\mu_c \geq 0$ and $\sigma^2 > 0$.

A model represents the set of *possible distributions*, indexed by **parameters**.

1. The "Family" of Candidates

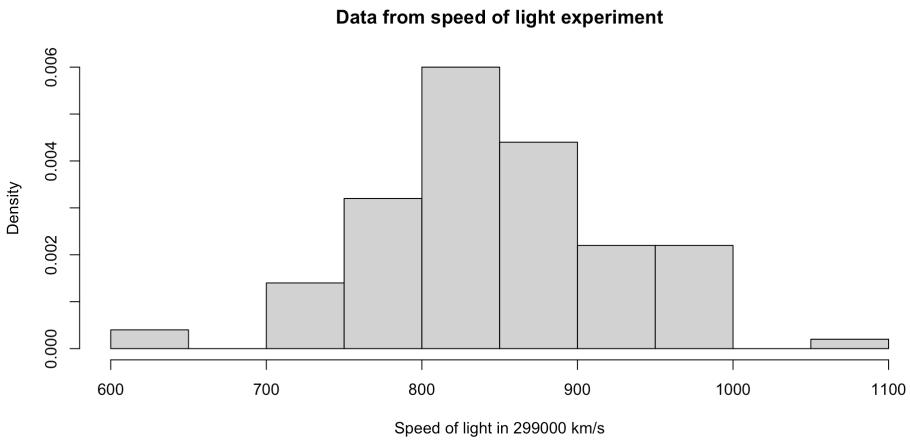
- The model suggests a family of distributions that could have generated our data.
- We don't know the "truth" yet, so we consider all plausible versions.

2. The Model Parameter (θ)

- Specifying the model **parameter(s)**, , picks one specific distribution out of the entire family.
- **Example**: For the speed of light, $\theta = (\mu_c, \sigma^2)$.
- Once we estimate a specific mean and variance, the model is fully specified.

Estimating parameters of the statistical model

Modelling example: the speed of light



Based on the model, how can we estimate the true speed of light $\mathbb{E}[X_i] = \mu_c$?

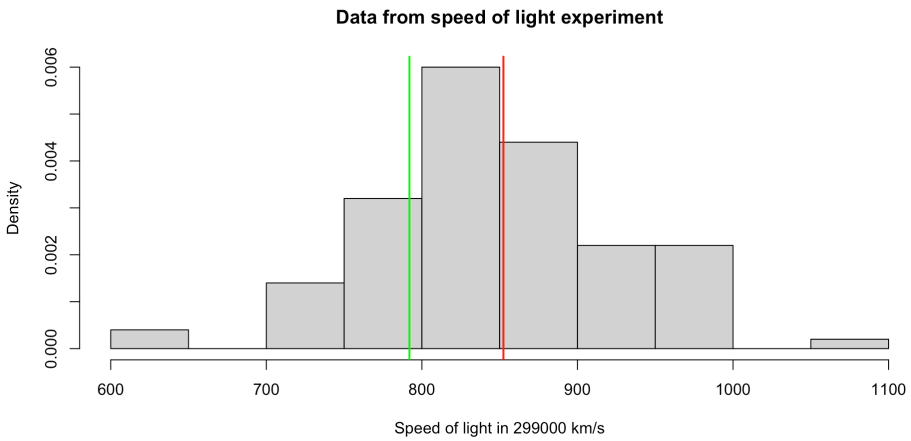
If our model is reasonable, we want to fin the expectation of the model distribution!

Chosen statistical model:

$$X_1, \ldots, X_{100} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_c, \sigma^2)$$

for $\mu_c \geq 0$ and $\sigma^2 > 0$.

Modelling example: the speed of light



An intuitive choice would be to use the **sample mean**.

- 299 852 km/s is the sample mean of the observations.
- 299 792 km/s is the actual speed of light in vacuum.

Estimators and estimates

- An **estimator** is a statistic we associate to a model parameter.
- Since it is a function of random variables, it is itself a random variable
- It is intended to approximate an unknown parameter of the data-generating distribution.

Example: is 5 an estimator for the speed of light μ_c?

Example: the sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is a function of the sample $X_1, \ldots, X_n$ and so is an estimator for μ_c.

Is the sample mean a *good* estimator?

- An **estimate** is a realization of an estimator associated to a model parameter.
- It is obtained *after* observing the data.
- For the sample mean, the estimate for the speed of light would be

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- Therefore, $\bar{x}_n$ would be the realization of the random variable $\bar{X}_n$
- In the speed of light example, the estimate using the sample mean would be $\hat{\mu}_c = \bar{x}_{100} = 299852$.

Estimators and estimates WLLN

If the model is correctly specified, then the WLLN tells us that

$$\bar{x}_n \xrightarrow{p} \mu_c.$$

This means that our estimate should get closer to the true speed of light as we increase our number of samples.

This means that $\bar{X}_n$ is a **consistent** estimator for μ_c.

Estimating the variability

What about the variability of measurements?

Suppose we have a random sample $X_1, \ldots, X_n$ and its mean μ is known but its variance is unknown. An estimator for the variance seen in class would be:

$$\tilde{S}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2.$$

Can we show that $\tilde{S}_n^2 \xrightarrow{p} \sigma^2$?

Summary

- We have gone through the process of statistical modelling for an example of experimental data
- We have used the example to motivate the definition of an **estimator**.