Introduction to computational statistics

LMS January 2026

* Descriptive statistics

* Foundations of probability

* Hypothesis testing

* Bayesian statistics

LMS Statistics and hypothesis testing

Chapter 1. Descriptive statistics

* Predictions and inference
* Populations and sampling
* True parameters and statistical estimators

Chapter 2. Foundations of probability

* Probability and random events
* Discrete probability ; Bernoulli, Binomial, Poisson, Uniform.
* Continuous probability ; Gaussian, Exponential, Uniform.

Chapter 3. Hypothesis testing (I)

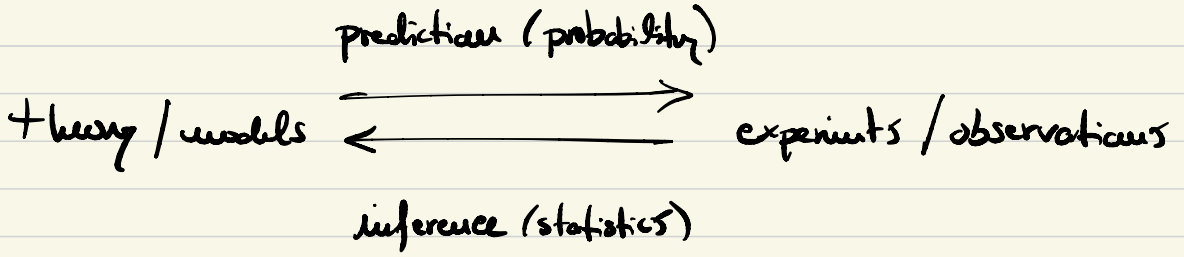
* the law of large numbers
* the central limit theorem
* Confidence intervals and critical regions.

Chapter 4. Hypothesis testing (II)

* the Fisher, Pearson, Neyman approach.
* Some examples : t-test, F-test, χ^2-test
* Parametric vs non parametric tests
* Error types in hypothesis testing (*) Bayesian statistics.

Introduction

* Prediction vs inference
* Probability, statistics and hypothesis testing
* Science works by first prediction and then experiment.
  Pearson, Fisher, Neyman (1920s) Bruno de Finetti (1974)



* Some sciences rely on predictive models and accurate predictions (e.g. Newtonian mechanics, planetary physics)
* Some sciences rely on inference and reconstruction "a posteriori" given some data (e.g. Darwinian evolution, historical geology)

i) Predictive probability;
   Cardano, Bernoulli, Laplace, Fermat (S. XVI / XVII)
ii) Descriptive statistics;
   Gauss, Nightingale, Pearson, Fisher (1900s)
iii) Hypothesis testing (*)
   Fisher, Pearson, Neyman (1930s)

1.1 Populations and samples

* Describing a large population of N individuals requires selecting / sampling a subset of $n < N$, from which we will make certain measurements

Population $P = \{x_1, \dots, x_N\}$

Sample $x = \{x_1, \dots, x_n\}$; $n < N$

* Every time I sample, my population becomes smaller. It is not the same to sample with and without replacement.

* Population is an idealized object, I will never have access to.

* I only have access to the observations of my finite sample

* How to ensure my sample accurately describes the population under study (*)

* Scale orders of magnitude

RNA seq ; 20.000 genes ; polyA tail capture + RT + PCR ;
10.000–15.000 ; 75 % of transcriptome

scRNA seq ; 20.000 genes ; single molecule capture (droplets) ;
Poisson sampling ; 5.000 ; 25 % of transcriptome

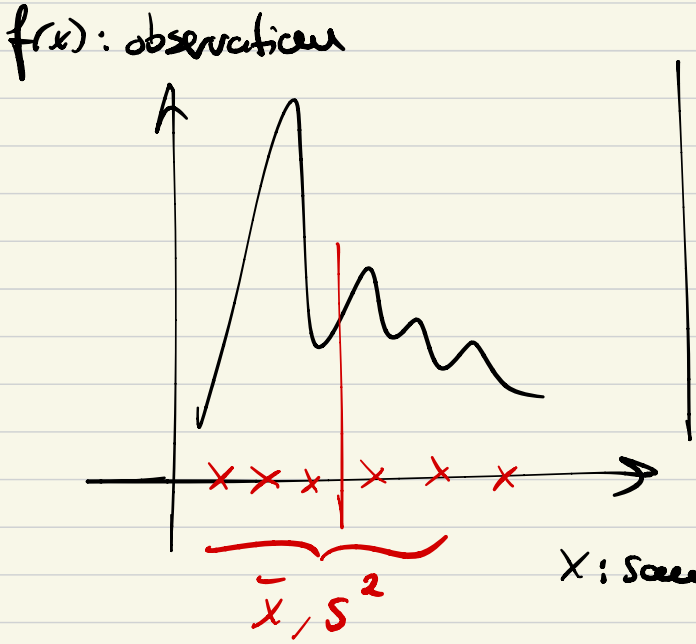
WGS ; $3 \cdot 10^9$ bp ; polymerase copying + optical base calling ;
$2.5–3 \cdot 10^9$ bp ; 95 % genome

DNA methylation ; 30M CpGs ; chemical / enzymatic C→T + seq.
15–20 M CpG sites ; 70 % CpG sites

Histone modifications ; $3 \cdot 10^9$ bp ; Antibody / epitope binding + seq.
20.000 – 100.000 peaks ; 1–5 % of genome,

1.2 Central tendency and variation

* Given a sample of n observations $x = \{x_1, \dots, x_n\}$
* We assume they are sampled from "same" distribution (μ, σ^2)



the sample mean $\bar{x}$ estimates
the central tendency
the sample variance s^2 estimates
the variation / spread within sample

* observed average ("sample mean")

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

* observed variance ("sample variance")

$$s^2 = \frac{1}{n-1} = \left\{ (x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \right\} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

↳ Bessel's factor (*)

{ Sample mean $\bar{x}$ is an estimator of μ
Sample variance s^2 is an estimator of s^2 }

Example : Compute sample mean and variance of

$x_1 = \{1, 2, 3\}$ and $x_2 = \{3, 4, 5\}$

Sample mean ; $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Sample variance ; $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

i) Sample mean and variance of x_1

$$\bar{x}_1 = \frac{1}{3}(1+2+3) = \frac{6}{3} = 2$$

$$s_1^2 = \frac{1}{2}\{(1-2)^2 + (2-2)^2 + (3-2)^2\} = 1$$

ii) Sample mean and variance of x_2

$$\bar{x}_2 = \frac{1}{3}(3+4+5) = \frac{12}{3} = 4$$

$$s_2^2 = \frac{1}{2}\{(3-4)^2 + (4-4)^2 + (5-4)^2\} = 1$$

1.3 Other measures of central tendency and variation.

* Imagine the sample $x = \{1, 4, 3, 2, 5, 37\}$

Compute sample mean

$$\bar{x} = \frac{1}{6}\sum_{i=1}^{6} x_i = \frac{1+4+3+2+5+37}{6} = \frac{52}{6} = \frac{26}{3} \simeq 8.7$$

Sample mean is sensitive to outliers; build a different estimator (*)

* Median

i) Sort elements in ascending order

$$x_{sorted} = \{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$$

ii) $$\text{med} = \begin{cases} x_{(\text{mid})} & \text{if } n \text{ odd } (n = 2k+1) \\ \dfrac{x_{(\text{mid}-1)} + x_{(\text{mid}+1)}}{2} & \text{if } n \text{ even } (n = 2k) \end{cases}$$

* Standard deviation

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

Example : Compute median $x = \{1, 4, 3, 2, 5, 37\}$

i) Sort ascending

$$x_{(sort)} = \{1, 2, 3, 4, 5, 37\}$$

ii) $n = 6$; even

$$med = \frac{x_{(3)} + x_{(4)}}{2} = 3.5$$ ✓

Example : $x = \{1, 4, 3, 2, 2, 5, 37\}$

i) Sort ascending

$$x_{(sort)} = \{1, 2, 2, 3, 4, 5, 37\}$$

ii) $n = 7$; odd

$$med = x_{(4)} = 3$$ ✓

Summary ; Descriptive statistics

* Prediction and inference

Probability (predictive) and statistics (descriptive)

* Population and Sample

$P = \{ x_1, \dots, x_N \}$ idealized, unaccessible.

$\eta = \{ x_1, \dots, x_n \}$ finite group / actually measure.

* Statistical estimators ; central tendency and variations.

i) Sample mean $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

ii) Sample variance $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

iii) Median med $= \begin{cases} x_{(\text{mid})} & \text{if } n \text{ odd } (n = 2k+1) \\ 1/2 \left(x_{(\text{mid}-1)} + x_{(\text{mid}+1)} \right) & \text{if } n \text{ even} \end{cases}$

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