Introduction to probability and statistics

Master in Cognitive Science 2025-2026

Lecture 5

November, 2025

Objectives

- Introduction to the basics of statistical inference (**frequentist** paradigm).
- Definition of an estimator.
- Important results: Law of Large Number (LLN) and the Central Limit Theorem (CLT).

Readings: Estimation Doc_fr4 file in Ametice. Kass et al Chapters 7.1, 7.2, 8.1

Definition:

- A population $\mathcal{P}$ is a large (sometimes infinite) set of individuals, objects, statistical units.
- A sample is a subset $\mathcal{E}$ of a population. It needs to be representative of the population:
  - the sample size n should be large if possible.
  - individuals randomly sampled from the population $\mathcal{P}$.

Link between probability and statistics

- *probability*: $\mathcal{P}$ is known, we compute the probability of observing some events in $\mathcal{E}$.
- *statistics*: $\mathcal{P}$ is unknown, we observe $\mathcal{E}$ and try to deduce properties of $\mathcal{P}$.

Example

Consider the following example: the population characteristic is the height of a person. Let X be a quantitative random variable representing a person's height. We can describe it as

$$X \sim \mathcal{N}(\mu = 175, \sigma^2 = 15^2)$$

.

```r
set.seed(123)
Pop = round(rnorm(10000, mean=175, sd = 15))
```

We pick randomly a sample $(x_1, \ldots, x_n)$ from the population.

```r
n = 10
samp1 <- sample(Pop)
samp1
```

In this case

```
## [1] 184 162 175 176 183 146 180 168 189 167
```

so if we compute the mean of *samp1*, we get an *estimation* of the mean of the population

```r
mean(samp1)
## [1] 173
```

Consider the dataset *bfi* from the package *psychTools*. We are interested on the age of the 2800 participants.

```
library(psychTools)
data(bfi)
help(bfi)
ages <- bfi$age
```

Visualization

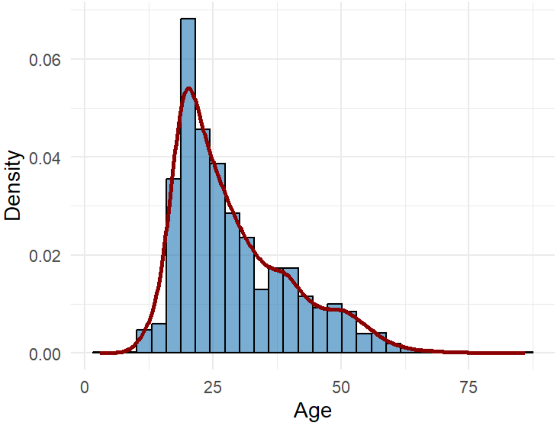

Figure 1: Histogram and Density of Age

Visualization of age against other variables

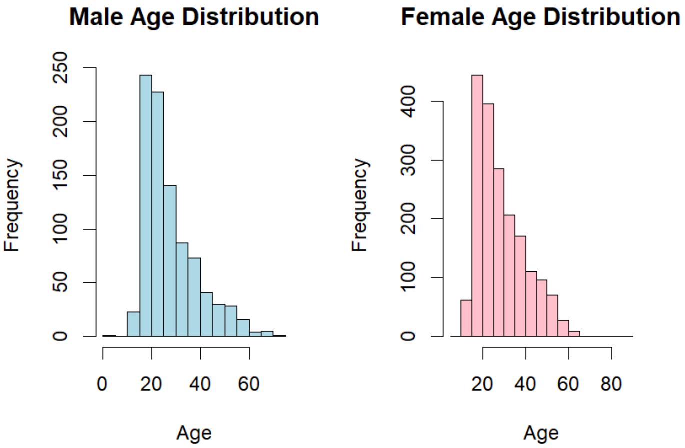

Figure 2: Comparison of age distributions between male and female

Measures of location

We could construct graphical description for our observed sample and/or compute numerical summaries:

- **The empirical mean**: $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.
- **Median**: we order our data $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ then take the value $x_{(\frac{n}{2})}$.

```
mean(ages)
## [1] 28.78214

median(ages)
## [1] 26
```

Boxplot

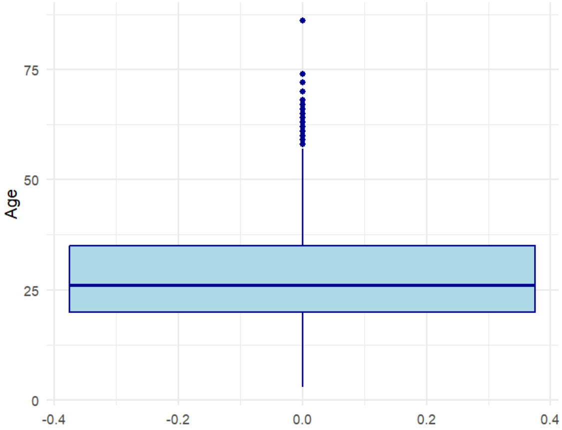

Figure 3: Boxplot of Participant Ages

Measure of spread

We use the variance to measure how far the data values lie from the mean. **The empirical variance**: $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.
Often, we use the corrected empirical variance

$$s_c = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

```
var(ages)
## [1] 123.8225

sd(ages)
## [1] 11.12755
```

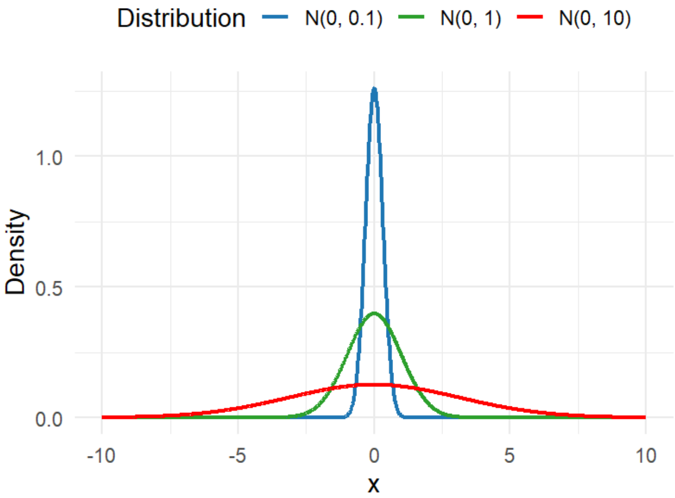

Figure 4: Normal Distributions with Different Variances

- We want to model the number of spike train of neuron 1 of *CAL1V* dataset in time interval 5 and 6. Suppose the number is a random variable $X \sim \mathcal{P}(\theta)$ for some θ.
- In this example there are 20 trials, so we model the number of spikes of each trial i by $X_i \sim \mathcal{P}(\theta)$.
- The trials are independent, so $X_1, X_2, \cdots, X_{20}$ are **independent**. Also they have the **same** distribution; in this case we say that they are **identically distributed**.
- We write shortly $X_1, X_2 \cdots, X_{20}$ are **i.i.d** to say that the random variables are independent and identically distributed.

1. From this data we get a realization of $X_1, X_2, \cdots, X_{20}$, which we denote by $x_1, x_2, \cdots, x_{20}$.

```
[1] 51 39 40 70 68 36 44 61 46 61 63 62 49 44
    44 51 64 55 47 58
```

2. An **estimator** of the parameter θ is given by

$$T(X_1, X_2, \cdots, X_{20}) = \frac{X_1 + X_2 + \cdots + X_{20}}{20}.$$

3. An estimation of the parameter θ is given by

$$\hat{\theta} = \frac{x_1 + x_2 + \cdots + x_{20}}{20} = 52.65$$

Estimator

Let $X_1, \cdots, X_n$ be an i.i.d. It forms a random sample.
The observed sample $x_1, \cdots, x_n$ is viewed as a realization of the random sample.
Suppose the distribution of X_i depends on a parameter θ.

Definition

An estimator of a population characteristic θ is a 'function' $T_n := T(X_1, \cdots, X_n)$ which estimates θ.

We need quality criteria to distinguish estimators. We say that T_n is

- unbiased if $E[T_n] = \theta$,
- efficient if $\mathrm{Var}(T_n)$ is small,
- consistent if $E[T_n] = \theta$ and $\mathrm{Var}(T_n) \to 0, n \to \infty$.
- The mean squared error of T_n is defined as $R(T_n, \theta) = E[(T_n - \theta)^2]$.

Example

Consider the estimator T_{20} of the average number of spikes.

Exercise

- Check that the estimator T_{20} is unbiased.
- Compute the variance of T_{20}.

Bias variance decomposition

Let T_n be an estimator of a parameter θ, then we have

$$\begin{aligned}
R(T_n, \theta) &= E\left[(T_n - \theta)^2\right] \\
&= E\left[(T_n - E[T_n] + E[T_n] - \theta)^2\right] \\
&= E\left[(T_n - E[T_n])^2\right] + (E[T_n] - \theta) \\
&= \mathrm{Var}(T_n) + \mathrm{Bias}^2(T_n).
\end{aligned}$$

We say that we decomposed the mean squared error into the variance of the estimator and the square of its bias.

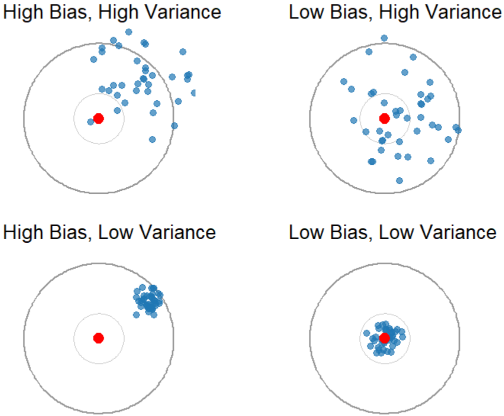

Figure 5: Bias–Variance Trade-off

Large Law Number

Theorem

Let $X_1, \ldots, X_n$ be i.i.d r.v. (independent identically distributed random variables) with $\mathbb{E}X_i = \mu$ and $Var(X_i) = \sigma^2$. Then,

$$\bar{X}_n := \frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu, \text{ as } n \to +\infty.$$

We say the random mean converges to μ (this holds in a sense that we won't discuss here).

Central Limit Theorem

Theorem

Let $X_1, \ldots, X_n$ be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then, as $n \to \infty$,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to \mathcal{N}(0, 1).$$

That is, the standardized sample mean converges to the standard normal law.