Mathematics for Data Science

Lecture 2bis: Introduction to Statistical Inference, Sampling and Limit Theorems

Mohamad GHASSANY

EFREI PARIS

Introduction to Statistical Inference

- **Statistics** is the science of collecting, processing and analyzing data derived from the observation of random phenomena.
- Data analysis is used to **describe** the phenomena studied, **make predictions** and **make decisions** about them. In this way, statistics is an essential tool for understanding and managing complex phenomena.
- The data studied can be of any nature, which makes statistics useful in all disciplinary fields.

The fundamental point is that the data present uncertainties and **variations**.

Statistical methods are divided into two classes:

- **Descriptive statistics**, **exploratory statistics** or **data analysis**, aims to summarize the information contained in the data in a synthetic and efficient way. Probabilities play only a minor role here.
- **Inferential statistics** goes beyond the simple description of data. Its purpose is to **make predictions** and **make decisions** based on observations. In general, it is necessary to propose **probabilistic models** of the studied random phenomenon and to know how to manage the risks of errors. Probabilities play a fundamental role here.

- **Probability** can be considered as a branch of pure mathematics, based on the theory of measurement, abstract and completely disconnected from reality.
- **Applied probability** proposes **probabilistic models** of the course of concrete random phenomena. One can then, **prior to any experiment**, make predictions about what will happen.

Example: it is usual to model the duration of the good functioning or life of a system, let's say a light bulb, by a random variable X of exponential law of parameter λ. Having adopted this probabilistic model, we can perform all the calculations we want. For example:

- The probability that the bulb has not yet failed at date t is $P(X > t) = e^{-\lambda t}$.
- The average lifetime is $E(X) = 1/\lambda$.
- If n identical light bulbs are turned on at the same time, and they work independently of each other, the number N_t of light bulbs that will fail before a time t is a random variable of binomial distribution $\mathcal{B}(n, P(X \leqslant t)) = \mathcal{B}(n, 1 - e^{-\lambda t})$. Thus we expect that, on average, $E(N_t) = n(1 - e^{-\lambda t})$ bulbs will fail between 0 and t.

In practice, if we want to use the theoretical results stated above, we have to make sure that we have chosen a good model, i.e.that the life span of these bulbs is a random variable with an exponential law, and, on the other hand, we have to be able to calculate the value of the parameter λ in some way. It is statistics that will allow us to solve these problems. To do this, we need to do an **experiment**, **collect data** and **analyze** them.

We therefore set up what we call a **test** or an **experiment**. We run $n = 10$ identical bulbs in parallel and independently of each other, under the same experimental conditions, and we record their lifetimes. Let's say that we obtain the following lifetimes, expressed in hours: $91.6, 35.7, 251.3, 24.3, 5.4, 67.3, 170.9, 9.5, 118.4, 57.1$

Let us note $x_1, \ldots, x_n$ these observations. We will therefore consider that $x_1, \ldots, x_n$ are the ***samples*** of random variables $X_1, \ldots, X_n$.

This means that after the experiment, the lifetime has been observed. We say that x_i is a sample (a realization) of X_i on the test performed.

Since the bulbs are identical, it is natural to suppose that X_i have the same law. This means that the same random phenomenon is observed several times.

We can also assume that the X_i are independent random variables. We can then ask the following questions:

The statistical approach iii

1. With respect to these observations, is it reasonable to assume that the lifetime of a light bulb is a random variable with an exponential distribution? If not, what other law would be more appropriate? This is a **fit test** (Chi-square test) problem.
2. If the exponential distribution model has been chosen, how can we propose a good value (or set of values) for the parameter λ? This is a parametric **estimation** problem.
3. In this case, can we guarantee that λ is less than a fixed value λ_0? This will guarantee that $\mathbb{E}(X) = 1/\lambda \geqslant 1/\lambda_0$, in other words that the bulbs will be sufficiently reliable. This is a **parametric hypothesis testing** problem.
4. If we have 100 light bulbs, how many failures can we expect in less than 50 hours? This is a **prediction** problem.

Sampling

Random Sample

The random variables $X_1, X_2, \ldots, X_n$ are a **random sample** of size n if (a) the X_i's are independent random variables, and (b) every X_i has the same probability distribution.

The observed data are also referred to as a random sample, but the use of the same phrase should not cause any confusion.

Statistic

A **statistic** is any function of the observations in a random sample.

For example, if $X, X_2, \ldots, X_n$ is a random sample of size n, the **sample mean** X, the **sample variance** S^2, and the **sample standard deviation** S are statistics. Since a statistic is a random variable, it has a probability distribution.

Sampling Distribution

The probability distribution of a statistic is called a **sampling distribution.**

For example, the probability distribution of X is called the **sampling distribution of the mean.**

The statistic $\overline{X}_n$

Consider determining the sampling distribution of the sample mean $\overline{X}$. Suppose that a random sample of size n is taken from a normal population with mean μ and variance σ^2. Now each observation in this sample, say, $X_1, X_2, \ldots, X_n$, is a normally and independently distributed random variable with mean μ and variance σ^2. Then, because linear functions of independent, normally distributed random variables are also normally distributed (Chapter 5), we conclude that the sample mean

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

has a normal distribution with mean

$$\mu_{\overline{X}} = \frac{\mu + \mu + \cdots + \mu}{n} = \mu$$

and variance

$$\sigma^2_{\overline{X}} = \frac{\sigma^2 + \sigma^2 + \cdots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

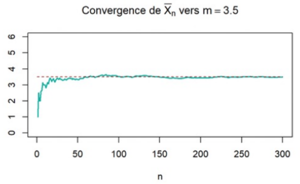
Limit Theorems

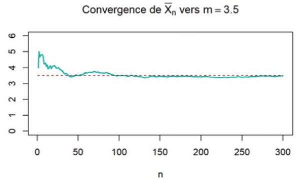
Strong law of large numbers

Let $X_1, X_2, \ldots, X_n$ be a set of independent random variables having a common distribution, and let $E[X_i] = \mu$. then, with probability 1

$$\frac{X_1 + X_1 + \ldots + X_n}{n} \to \mu \quad \text{as } n \to \infty.$$

x_i	1	2	3	4	5	6
$P(X = x_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$





Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population (either finite or infinite) with mean μ and finite variance σ^2, and if $\overline{X}$ is the sample mean, the limiting form of the distribution of

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \qquad (7\text{-}1)$$

as $n \to \infty$, is the standard normal distribution.

Application 1: For a big enough sample size n, we can consider that $\overline{X}_n$ has as distribution:

$$\overline{X}_n \sim \mathcal{N}(m, \frac{\sigma^2}{n})$$

Application 2: The distribution of a percentage, studied in the next section.

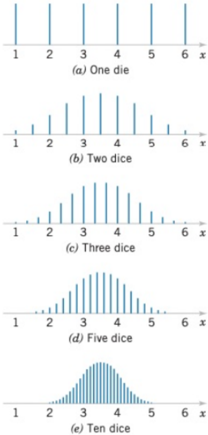


Figure 7-1 Distributions of average scores from throwing dice. [Adapted with permission from Box, Hunter, and Hunter (1978).]

The statistic P_n (percentage)

Let X be the random variable representing the number of successes in a series of n independent repetitions of the same test with probability p.

The distribution of X is the binomial distribution of parameters n and p, denoted $\mathcal{B}(n, p)$. X is the sum of n independent Bernoulli variables of parameter p.

Let P_n be the *empirical frequency* of the number of successes among the n trials: $P_n = \frac{X}{n}$

$P_n = \overline{X}_n$ because X is the sum of n independent Bernoulli variables of parameter p.

P_n has expectation and variance:

$$E(P_n) = p \quad \text{and} \qquad V(P_n) = \frac{p(1-p)}{n}$$

Applying the central limit theorem to X sum of Bernoulli variables:

For n sufficiently large, we can consider that P_n follows the normal distribution:

$$P_n \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

Empirical Variance

The S_n^2 statistic or empirical sample variance is defined by:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Properties:

- $S_n^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - (\overline{X}_n)^2$.
- $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - (\overline{X}_n - m)^2$.
- S_n^2 almost surely converges to σ^2.

Expectation S_n^2 **is**:

$$E(S_n^2) = \frac{n-1}{n}\sigma^2$$

demonstration:

$$\begin{aligned} E(S_n^2) &= \frac{1}{n}\sum_{i=1}^{n} E(X_i - m)^2 - E(\overline{X}_n - m)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} V(X_i) - V(\overline{X}_n) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 - \frac{\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \end{aligned}$$

Remark that if we let: ${S_n^*}^2 = \frac{n}{n-1}S_n^2$ so $E({S_n^*}^2) = \sigma^2$.