Inferential Statistics
L2 - Descriptive statistics and statistical models

Erlis Ruli (erlis.ruli@unipd.it)

Department of Statistics, University of Padova

Contents

The samples

The data collected in an experiment consist of observations $x_1, x_2, \ldots, x_n$ on a variable of interest, which are then used to learn about the data-generating mechanism.

The list $x_1, \ldots, x_n$ is called the observed sample and n is called the sample size.

We assume that $x_1, \ldots, x_n$ is a realisation of the random sample $X_1, \ldots, X_n$, with X_i assumed mutually independent with equal marginal pdf f.

The distinction between observed and random sample is much like the difference between a measured voltage (observed) and the voltmeter.

By the definition of independence, the joint pdf of the sample is

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta),$$

where $f(x_i; \theta)$ is the density for X_i which depends on some unknown parameter θ.

Example 1

Let $X_1, \ldots, X_n$ be a random sample from the population $\text{Exp}(1/\beta)$. X_i may be time (years) until failure for n identical circuit boards put to test.
The joint pdf is

$$f(x_1, \ldots, x_n; \beta) = \prod_{i=1}^{n} f(x_i; \beta) = \prod_{i=1}^{n} \tfrac{1}{\beta} e^{-x_i/\beta} = \tfrac{1}{\beta^n} e^{-(x_1 + \cdots + x_n)/\beta}$$

We could use this to compute, say the probability that all boards last at least 5 years:

$$P(X_1 \geq 5, \ldots, X_n \geq 5) = \int_5^{\infty} \cdots \int_5^{\infty} \prod_{i=1}^{n} \tfrac{1}{\beta} e^(x_i/\beta) dx_1 \cdots dx_n = e^{-5n/\beta}.$$

Summary statistics

Typically we are interested at some function of the sample. These are called descriptive or summary statistics.

Some examples are moment-based statistics:

- sample average $\overline{X} = \frac{1}{n}\sum_i X_i$ and the observed counterpart $\overline{x} = \frac{1}{n}\sum_i x_i$
- sample variance $S^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$ and the observed counterpart $s^2 = \frac{1}{n-1}\sum_i (x_i - \overline{x})^2$; s is commonly called standard deviation.
- sample kth moment $\overline{X^k} = \frac{1}{n}\sum_i X_i^k$ and the observed counter part.

Order statistics

Let $X_{(1)} = \min_{1 \le i \le n} X_i$ be the smallest observation, $X_{(2)}$ be the second smallest and so on $X_{(n)} = \max_{1 \le i \le n} X_i$.

The list $X_{(1)}, \ldots, X_{(n)}$ is called order statistics, and are the basis of the following summary statistics

- the median $Q_2 = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ (X_{(n/2)} + X_{(n/2+1)})/2 & \text{if } n \text{ is even} \end{cases}$
- the first and third quartile, $Q_1 = X_{[0.25(n+1)]}$ and $Q_3 = X_{[0.75(n+1)]}$, resp
- the pth sample quantile, $p \in (0, 1)$ is $X_{[p(n+1)]}$
- inter quartile range IQR $= Q_3 - Q_1$
- median absolute deviation from the median (MAD) $=$ median$(|X_1 - Q_2|, \ldots, |X_n - Q_2|)$

and their observed counterparts; $[x]$ is the greatest integer $\le x$.

Uses and relations

The above summary statistics serve different purposes:

$\overline{X}$, Q_1, Q_2, Q_3, $X_{[p(n+1)]}$ are measures of location and are used when we want to provide a single typical value of the sample

S^2, S, MAD, IQR are measures of spread, useful when we want to describe the variability of the sample.

→ "non-simmetry"
→ length of the "tail"

You might heard about skewness, kurtosis. These are additional features of the shape of distribution/sample.

Sample measures target their population counterparts, e.g. $\overline{X}$ for μ_X, Q_2 for $\xi_0.5$ S^2, MAD for σ^2, etc.

Example 2

Suppose the sample of size is $n = 12$ and the 0.65th quantile is wanted. Then $[0.65 \cdot (12 + 1)] = 8$, so the 0.65th quantile is $X_{(8)}$. The answer would have been the same if wanted the $0.69th$ quantile.

(Two different examples) nice

Consider now the observed sample $1.1, 0.5, 0.4, 3, 2.2$, so $x_1 = 1.1$, $x_2 = 0.5$ and so on. The observed order statistics are

$$x_{(1)} = 0.4, x_{(2)} = 0.5, x_{(3)} = 1.1, x_{(4)} = 2.2, x_{(5)} = 3.$$

We find that $\overline{x} = 1.44$, $s^2 = 1.273$, $q_1 = 0.4$, $q_2 = 1.1$, $q_3 = 2.\cancel{3}2$, and mad$= 1.03782$.

Histogram

Useful when we want to get an idea of the pdf of a sample.

Let $x_1, \dots, x_n$ be the observed sample and consider a partition in intervals $(a_{j-1}, a_j]$, $j = 1, \dots, m$, with $m < n$ covering the sample.

Defined by the piecewise function

$$h_n(x) = \frac{1}{n(a_j - a_{j-1})} \sum_{i=1}^{n} \mathbf{1}_{(a_{j-1}, a_j]}(x_i), \quad \text{for all } x \in (a_{j-1}, a_j].$$

Typically, $(a_{j-1}, a_j]$ are equal-length intervals and $m = 2\,\mathrm{iqr}/n^{1/3}$ (Friedman-Diaconis rule). The $h_n(x)$ thus targets $f(x)$, the population pdf, i.e. the pdf from which the observations come from.

Empirical distribution function

converges faster than the histogram to the true pdf

Given the random sample (rs) $X_1, \ldots, X_n$, the edf is defined by

$$F_n(x) = \sum_{i=1}^{n} I_{X_i}(x), \quad \text{for all } x \in \mathbb{R},$$

where $I_{X_i}(x)$ is Bernoulli rv with success probability $P(X_i \leq x)$. For each x, F_n is thus a random variable.

The corresponding observed version is

$$\widehat{F}_n(x) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{x_i}(x), \quad \text{for all } x \in \mathbb{R},$$

$\mathbf{1}_{x_i}(x)$ takes value 1 if $x_i \leq x$ and 0 otherwise.

F_n and its observed version $\widehat{F}_n$ target $F(x)$, the population df.

Example 3

Compute $\widehat{F}_n$ from the observed sample $1.1, 0.5, 0.3, 1.1, 5$.

First, we have to get the sorted list, which is $0.4, 0.5, 1.1, 1.1, 5$. Then we observe that

- for $-\infty < x < 0.4$ there are no observations, so $\sum_i \mathbf{1}_{x_i}(x) = 0$
- for $0.4 \leq x < 0.5$ there is only one observation, so $\sum_i \mathbf{1}_{x_i}(x) = 1$
- and so on,
- for $1.1 \leq x < 5$ there are two observations, so $\sum_i \mathbf{1}_{x_i}(x) = 2$.

Hence

$$\widehat{F}_n(x) = \begin{cases} 0 & \text{if } x < 0.4 \\ 1/5 & \text{if } 0.4 \leq x < 0.5 \\ 2/5 & \text{if } 0.5 \leq x < 1.1 \\ 4/5 & \text{if } 1.1 \leq x < 5 \\ 1 & \text{if } 5 \leq x. \end{cases}$$