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 <u>order statistics</u>, 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 <u>measures of location</u> and are used when we want to provide a single typical value of the sample

$S^2, S, \text{MAD}, \text{IQR}$ are <u>measures of spread</u>, useful when we want to describe the variability of the sample.

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.

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 $\overline{x} = 1.44$, $s^2 = 1.273$, $q_1 = 0.4$, $q_2 = 1.1$, $q_3 = 2.3$, and mad$= 1.03782$.

Histogram

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

Let $x_1, \ldots, x_n$ be the observed sample and consider a partition in intervals $(a_{j-1}, a_j]$, $j = 1, \ldots, m$, 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

Given the random sample (rs) $X_1, \dots, X_n$, 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

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

Boxplot (Box-and-whiskers plot)

It's a 5-summary measures description of a (typically observed) sample.

It's provides informations about the location, spread and the shape of the distribution of the sample.

In the vertical orientation:

- the middle line represents q_2, and vertical edges of the box represent q_1 and q_2, resp.
- the upper whisker is largest $x_i \leq q_3 + 1.5 \cdot \text{iqr}$
- the lower whisker is the smallest $x_i \geq q_1 - 1.5 \cdot \text{iqr}$
- observations outside the whisker are typically marked by a "*"
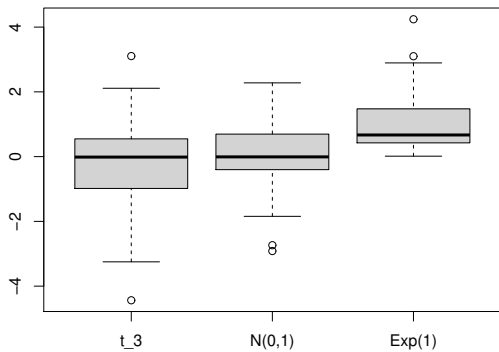
Quantile-Quantile plot

The QQ plot is useful for checking if an observed sample is compatible with a population with continuous F.

It works by comparing a list of observed quantiles of the sample with the corresponding quantiles of F.
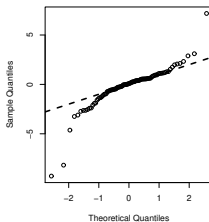
It consists in plotting the pairs $(x_{(i)}, F^{-1}(i/(n+1)))$ and looking for a linear relationship.

The QQ plot with F the normal distribution is the most widely used. In practice F involves unknown parameters which have to be estimated beforehand.
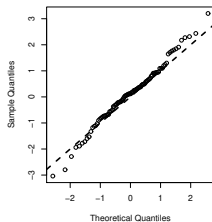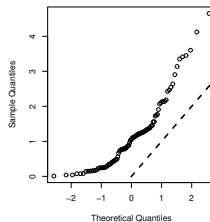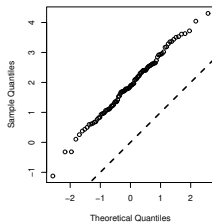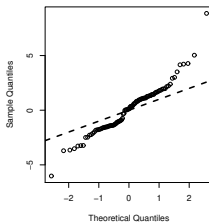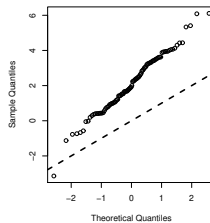
In realistic applications we may collect observation for several variables, thus a typical dataset looks like

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

where n is the number of observations and p is the number of variables.

The jth column represents the overall sample for the jth variable, and the ith row is the ith sample point for all variables.

For example, the columns could be pollutants s.t. $PM_{2.5}$, PM_{10}, CO_2, etc. and the rows may be values measured hourly.

Summaries for multivariate data

A common query is if the *p* variables are related to each other.

A first approach could be to plot pairs of variables and inspect the graph for possible associations.

For pairs of variables the sample covariance and the sample Pearson's correlation are widely used measure of association.
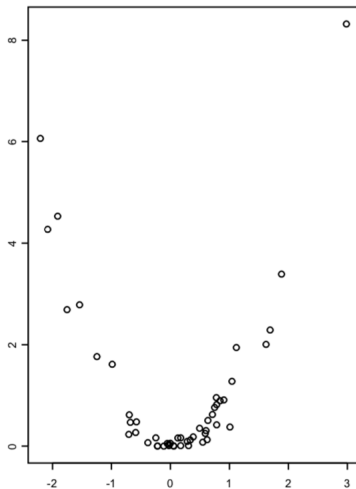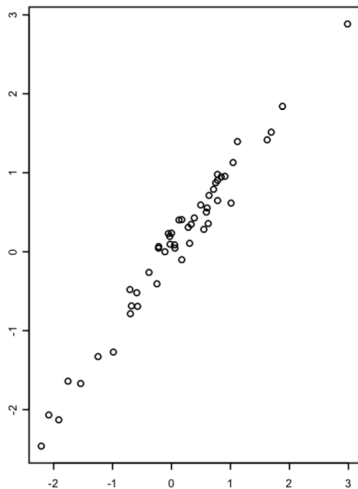In particular, for a pair of variables x, y, the sample covariance is

$$s_{xy} = (n-1)^{-1} \sum_i (x_i - \overline{x})(y_i - \overline{y}),$$

and the sample Pearson correlation is

$$r_{xy} = \frac{s_{xy}}{s_x \, s_y}.$$

s_{xy} targets it population version σ_{XY}, whereas r_{xy} targets its population version ρ_{XY}.

Caution: lack of correlation $\neq$ lack of association!

Let $X_1, \ldots, X_n$ be random sample with $X \sim F_\theta$. If, in addition, X_i are also independent we call it iid random sample.

The joint pdf of the sample is $f x_1, \ldots, x_n; \theta = \prod_{i=1}^{n} f(x_i; \theta)$.

By a <u>statistical model</u> we mean the set

$$\{ f(x_1, \ldots, x_n; \theta) : \theta \in \Theta, x_i \in \mathcal{X} \},$$

where Θ is the set of all possible values for θ.

Typically, $\Theta \subseteq \mathbb{R}^d$ for some integer $d > 0$ and X_i could be a rv or a rve of any dimension.

Example 4

Let $X_1, \ldots, X_n$ be an iid random sample with $X_i \sim \text{Poi}(\lambda)$. The joint distribution of the sample is

$$f(x_1, \ldots, x_n; \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}}{x_i!} \lambda^{x_i}$$

$$= e^{-n\lambda} (\lambda)^{\sum_i x_i} \left(\prod_i x_i! \right)^{-1},$$

and the statistical model is

$$\left\{ e^{-n\lambda} (\lambda)^{\sum_i x_i} \left(\prod_i x_i! \right)^{-1} : \lambda > 0, x_i \in \mathbb{N}, \text{for all } i \right\}.$$

Statistical models for non iid data

In many applications the iid assumption, especially the "identically" part, is unrealistic. Here is a common situation.

P1 produces washing machine (WM) motors. He claims that his new model (NM) is more efficient, while achieving the same speed as the old, version motors (OM). Which one should we buy? To answer this question, we have to run experiments with WM+NM and WM+OM and then analyse the data. Which statistical model should we use for this problem?

Two-sample normal model

Let the $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be the energy consumptions measured under the OM and NM respectively.

It's reasonable to assume that

- the measures within each motor independent (discussion?)
- the measures between motors are also independent
- the data-generating process under OM may differ from that of NM
- energy consumption is reasonably Gaussian (sum of many small energetically hungry components), thus

Joint distribution for OM: $f(x_1, \ldots, x_m, \theta_x) = \prod_{i=1}^{m} f(x_i; \theta_x)$

Joint distribution for NM: $f(y_1, \ldots, y_n) = \prod_{j=1}^{n} f(y_j; \theta_y)$.

Joint distribution for OM and NM:

$$f(x_1, \ldots, x_m, y_1, \ldots, y_n; \theta_x, \theta_y) = \left(\prod_i f(x_i; \theta_x) \right) \left(\prod_j f(y_j; \theta_y) \right),$$

where $\theta_x = (\mu_x, \sigma_x^2), \theta_y = (\mu_y, \sigma_y^2)$. The statistical model is the set of all joint distributions generated by the different parameter values.

Linear regression

Suppose we have a device able to remove arsenic (which has negative health effects on human beings) from drinkable water and we suspect the effectiveness of the removal depends on the pH of water. Se we need to assess how arsenic removal is affected by water pH.

Let $x_1, \ldots, x_n$ be pH values of n samples of water and let $y_1, \ldots, y_n$ be the measured values of arsenic removed from each of the water samples.

Typically we assume $x_1, \ldots, x_n$ are fixed and the y_i are a realisation of the random sample $Y_1, \ldots, Y_n$.

Then, a possible model is:

$$Y_i | x_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots$$
$$Y_i \text{ independent from } Y_j \quad \text{for all } i, j.$$

The unknown parameter is $(\beta_0, \beta_1, \sigma^2)$.

Logistic regression

Suppose you want to predict chicken sex from its egg features.

For the ith egg, let $x_{i1}, \ldots, x_{ip}$ be p features (e.g. volume, color, etc.) and let $y_1, \ldots, y_n$ be the chicken sex (1=female, 0=male). A simple model for this problem can be built as follows:

$$Y_i | x_{i1}, \ldots, x_{ip} \sim \text{Ber}(\theta_i)$$
$$\theta_i = \frac{1}{1+e^{-\mu_i}}$$
$$\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$
$$Y_i, Y_j \quad \text{independent for all } i, j.$$

The joint distribution is

$$f(y_1, \ldots, y_n | \mathbf{X}, \theta) = \prod_{i=1}^{n} \left(\frac{1}{1+e^{-(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}} \right)^{y_i} \left(\frac{1}{1+e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}} \right)^{1-y_i}.$$

The statistical model is the set of joint distributions at all $\theta = (\beta_0, \ldots, \beta_1) \in \mathbb{R}^{p+1}$.