Data, Models, Parameters, and Statistics

February 13, 2025

Motto

Nothing is more practical than a good theory.

Vladimir Vapnik*

*in *Statistical Learning Theory*. John Wiley, New York (1998)

Data as an outcome of sampling variables

1. Empirical observations are often referred to as *data*.
2. Thus a set of anything (numbers, categorical values, names, good, bad, thick, thin, woman, man, images or even movies, etc.) that has been obtained or extracted from empirical evidence can be called a data set
3. In the theory of statistics, data are characterized by two fundamental aspects:
   - They can be grouped according to a particular structure relevant to the purpose of data collection, such a grouping is called a ***data point*** or a ***datum***. If there is only one value per a data point, then one deals with one-dimensional data point (no structure within a data point).
   - Typically, there are many data points in a data set and each individual data point has similar (if not identical) structure as other data points (belongs to the same domain).
4. Data points form a ***sample*** and there is another nature of the structure between data points than within data points .

Structure of a datum

Data are mathematical representations of observations. We roll a die, we see the number of dots that came up, say six, and we say that our datum (data point) is 6. Data can consist of

1. Vectors of scalars, measurements, and/or characters, for example, a time series of measurements.
2. Matrices of scalars and/or characters, for example, digitized pictures or more routinely measurements of covariates and response on a set of individuals—see Example 1.1.4 and Sections 2.2.1 and 6.1.
3. Arrays of scalars and/or characters as in contingency tables—see Chapter 6—or more generally multifactor-multiresponse data on a number of individuals.
4. Mixtures of all of the above and more, in particular, functions as in signal processing, trees as in evolutionary phylogenies, and so on.

Random character of sampling

In statistics, almost invariably, we observe many data points by observing them repetitively. The repetitive collection of structurally the same (or similar) data points is referred to as ***sampling***.

1. The most classical and elementary sampling is ***independent sampling*** under the same circumstances.
2. Sampling under different circumstances affected by some covariates (different geographical locations or by different methods of data collection methods, etc.)
3. Dependent sampling can occur as well (for example, in time series).
4. The most important about sampling is that its mechanism is well understood and follow the theoretical paradigms that we assume about the model.

The number of repetitive data points is referred to as a ***sample size***.

Statistical model – clarification of randomness

A statistical model is mathematical conceptualization of the stochastic mechanism that produces the data.

- $\mathcal{X}$ – **sample space**, the set all possible values for data.
- $(\Omega, \mathcal{F}, P)$ – a probability space, where Ω is a set elementary outcomes, $\mathcal{F}$ is a σ-field of subsets of $\mathcal{F}$, a.k.a. a collection of events, P is a probability measure on $\mathcal{F}$.
- Random variable X is a mapping from Ω to $\mathcal{X}$ (more precisely a measurable mapping).
- Our data are resulting from observing $X(\omega)$, where the choice of ω is dictated by the probability measure P.
- If P is known, then there is no statistics only probability theory.
- Thus statistics (as a discipline) deals with the cases when P is unspecified, for example, we only know that it belongs to a certain family of probability measures

$$\{P_\theta\}_{\theta\in\Theta}.$$

Statistical Model on space of observations

One can simplify the set-up and consider only the probabilities on the space of values of X, i.e. on $\mathcal{X}$. This is because any P_θ generates distribution of X on events in $\mathcal{X}$, i.e. for an event $A \subset \mathcal{X}$

$$P_\theta^X(A) = P_\theta(X \in A) \left(= P_\theta(X^{-1}(A))\right)$$

- Thus a statistical model deals with parametrized probability measures on the space of observations (sample space) and we drop X from P_θ^X:

$$A \subset \mathcal{X} \mapsto P_\theta(A),\ \ \theta \in \Theta.$$

- The true θ is unknown
- The goal of statistics is to find out θ from what was observed i.e. from the observed $x = X(\omega)$, i.e. from the data.

Example - hypergeometric model

We are faced with a shipment of N manufactured items. An unknown proportion θ of these elements are defective. It is too expensive to examine all of the items. So to get information about θ, a sample of n elements is drawn without replacement and inspected. The data is made of one observation that is the number k of defectives found in the sample.

- Data (observation) x. Strictly speaking we can have either n data points (0-defective or 1-good) or only one datum (the number of defectives).
- The random variable X is a number of defective in a random experiment of drawing n from N.
- The statistical model

$$P_\theta(X = k) = \frac{\binom{N\theta}{k}\binom{N-N\theta}{n-k}}{\binom{N}{n}}.$$

Example - sampling population

Suppose that in a small community, we approach individuals that are eligible for voting and we are interested in finding if they are planning to take part in the coming election. To review the introduced basic concepts, let us answer the following questions, in the relation to the hyperbolic model

- What is a datum?
- What is a sample?
- What is a sample size?
- What kind of sampling do we deal with?
- What a statistical model is suitable for the problem?
- What is (are) a parameter(s) in this problem?

Homework *Propose a hierarchical extension of this model to capture the percentage of voting population in a country that is treated a conglomerate of small communities.*

Regular models

Notation:

- θ: a parameter specifying a probability distribution P_θ.
- $F(\cdot|\theta)$: Distribution function of P_θ
- $E_\theta[\cdot]$: Expectation under P_θ. For a (measurable) function $g(x)$:

$$E_\theta[g(X)] = \int_{\mathcal{X}} g(x) dF(x|\theta)$$

- $p(x|\theta) = p(x;\theta)$: probability-density or -mass function (pdf or pdm) of P_θ
- **Regularity assumptions**:
  - **Either** All P_θ's are (absolutely) continuous with densities $p(x|\theta)$,
  - **Or** All P_θ's are discrete with pmf's $p(x|\theta)$ and the set $\{x : p(x|\theta) > 0\}$ is the same for all $\theta \in \Theta$, i.e. the common support of distributions.

Parameter space

- A statistical model is essentially described by a family of probabilities that is feasible for a given space of observations.
- Such a family can be always parametrized $\{P_\theta\}_{\theta \in \Theta}$.
- Actual observations (data) come from the model P_{θ_0} and $\theta_0 \in \Theta$ is called a true (while unknown) parameter.
- It is typically assumed that θ are fully identifiable, i.e. different parameters lead to different probabilities (distributions).
- If Θ is a subset of a finite dimensional space, the model is called ***parametric***.
- If Θ is a part of infinite dimensional space, the model is called ***non-parameteric***.
- If some natural sub-parameters in Θ belong to finite dimensional space and this part is of the main interest while the rest is infinite-dimensional but not of the interest model is called ***semi-parametrics***.

Example

We want to study how a physical or economic feature, for example, height or income, is distributed in a large population. An exhaustive census is impossible so the study is based on measurements and a sample of n individuals drawn at random from the population. The population is so large that the actual process can be considered as sampling with replacement. For example, an experimenter makes n independent determinations of the value of a physical constant μ. His measurements are subject to random fluctuations (error) and the data can be thought of as μ plus some random errors.

- The model:

$$X = (X_1, \ldots, X_n) = (\mu + \epsilon_1, \ldots, \mu + \epsilon_n),$$

  where ϵ_i are independent identically distributed according to a symmetric-around-zero distribution G or, for shortness,

$$\epsilon_i \overset{iid}{\sim} G$$

Parameter space in the example

- The data $X \in \mathbb{R}^n$ have the distribution defined by the cumulative distribution function (cdf)

$$P(X \leq x) = \prod_{i=1}^{n} G(x_i - \mu)$$

- $\theta = (\mu, G)$ and the parameter space is $\mathbb{R} \times \mathcal{P}_0$, where $\mathcal{P}_0$ is the class of all symmetric around zero distributions.
- The model is non-parametric because θ is not from a finite dimensional space ($\mathcal{P}_0$ is not a finite dimensional space).
- However, μ is typically the parameter of interest and G is only a nuisance parameter. Since μ is one-dimensional and of interest, this case is defined as a semi-parametric case.
- Finally, we can limit the class of possible G for the model by, for example, assuming that $G = N(0, \sigma^2)$. The model becomes fully parametric with $\theta = (\mu, \sigma^2) \subset \mathbb{R} \times (0, \infty)$.

A statistic

- A ***statistic*** is a (measurable) function of an observation that *does not depend* on the specific model generating a given type of data.
- Assume that the data are observation from the model given by $X \in \mathcal{X}$ and with the distribution from the class $\{P_\theta\}_{\theta \in \Theta}$.
- Any (measurable) function $T : \mathcal{X} \mapsto \mathcal{T}$, when appled to the random variable X is called a statistics, i.e. $T(X)$ is called a statistics.
- The two most known statistics are sample mean and sample variance, i.e. for $X = (X_1, \dots, X_n)$:

$$T_1(X) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$T_2(X) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - T_1(X))^2$$

In the traditional notation $\bar{X} = T_1(X)$ and $S^2 = T_2(X)$.

More examples

The statistic evaluated in a concrete situation typically aims at some statistical parameter (the mean, variance).

- The proportion of defective elements X/n in a sample of manufactured elements is also a statistic.
- The empirical distribution function $\widehat{F}_n(X)$ aims at the (true) cdf F of X_i's

$$\widehat{F}_n(X)(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(0,x]}(X_i) = \frac{\#\{i \le n : X_i \le x\}}{n}.$$

Of course, $\widehat{F}_n(X)$ is also a statistic.

Homework *What is the pointwise limit of $F_n(X)$, when n increases without bound? Use the Law of Large Numbers. Use the Central Limit Theorem to establish the pointwise bounds for the error $F_n(X) - F$. Can you elaborate why the word 'pointwise' is used here?*

Statistical inference

A statistic should be viewed as a certain summary of some information contained in the data. It serves of a certain purpose to conclude something about the underlying statistical model, or as we say, to make **statistical inference**. The following are traditional goals of such inference

- **Estimation** – producing “best guesses” of the values of important model parameters – **point estimation**.
- **Confidence intervals** – assessing the error of a point estimator – **interval estimation**.
- **Testing significance** – determining if the data support a certain specific claim about the model – **testing statistical hypotheses**.
- **Prediction** – finding the best method of predicting an unknown variable based on the observed ones and assessing the prediction error.

Decision theory framework

Sometimes it is convenient to formulate the statistical inference problem in the framework of the decision theory.

- Based on the data x a statistician makes a decision $\delta(x)$ about some function $\nu(P_{\theta_0})$ of the true distribution given by P_{θ_0}.
- There is some **loss** associated with this decision defined by $l(\theta_0, \delta(x))$.
- It can be, for example, the **quadratic loss** if the inference is about the parameter $\nu_0 = \nu(P_{\theta_0})$, so $\delta(x) = \hat{\nu}(x)$ is a value of an estimator, then
$$l(\theta_0, \delta(x)) = (\nu_0 - \hat{\nu}(x))^2$$
- The **risk** function is the mean loss (since it is averaged, it is not depending on the data anymore)
$$R(\theta_0, \delta) = E_{\theta_0} l(\theta_0, \delta(X))$$
- In the quadratic loss, the risk becomes the mean square error (MSE)
$$R(\theta_0, \delta) = E_{\theta_0}(\nu_0 - \hat{\nu}(X))^2$$

The MSE, variance, and bias

- The mean square error (MSE) satisfies

$$E_{\theta_0}(\nu_0 - \hat{\nu}(X))^2 = Var(\hat{\nu}) + Bias^2(\hat{\nu}),$$

  where $Bias(\hat{\nu}) = E_{\theta_0}(\hat{\nu}) - \nu_0$.
- We often drop θ_0 from the notation if it is clear that the expectations are taken with respect to the true distribution standing behind the data.

The following is to remember

MSE=Variance+Bias^2

General regression models

- We observe $(z_1, Y_1), \ldots, (z_n, Y_n)$, where $Y = (Y_1, ..., Y_n)$ is a vector of independent variables.
- The distribution of the response Y_i for the *i*th subject or case in the study is postulated to depend on certain characteristics z_i of the *i*th subject. Thus, z_i is a d dimensional vector, aka a covariate or an explanatory variable, that gives characteristics such as sex, age, height, weight, and so on of the *i*th subject in a study.
- If we let $f(y_i|z_i)$ denote the density of Y_i for a subject with covariate vector z_i, then the full statistical model for Y is given by the joint density

$$p(y) = \prod_{i=1}^{n} f(y_i|z_i)$$

Specifications of the regression model

- A special non-parameteric subcase, a function μ is unknown and

  $$Y_i = \mu(z_i) + \epsilon_i,\ i = 1, \dots, n.$$

  where $\epsilon_i \overset{iid}{\sim} F$.

- A special subsubcase, a function g is known but β is unknown

  $$\mu(z) = g(\beta, z),$$

  g known, $\beta \in \mathbb{R}^d$ unknown.

- A special subsubsubcase (linear model):

  $$g(\beta, z) = z^\top \beta$$

- A special subsubsubsubcase, a parametric model (classical Gaussian model):

  $$\epsilon_i \sim N(0, \sigma^2).$$

Autoregressive errors

Let $X = (X_1, ..., X_n)$ be the n determinations of a physical constant μ.

$$X_i = \mu + e_i, \quad i = 1, \ldots, n$$
$$e_i = \beta e_{i-1} + \epsilon_i, \quad i = 1, \ldots, n, \quad e_0 = 0$$

where $\epsilon_i \overset{iid}{\sim} F$.
An example would be, say, the elapsed times $X_1, ..., X_n$ spent above a fixed high level for a series of n consecutive wave records at a point on the seashore. Let $\mu = E(X_i)$ be the average time for an infinite series of records. It is plausible that e_i depends on e_{i-1} because long waves tend to be followed by long waves.

Homework *Based on your empirical knownledge of wave sizes on a stormy day in Malmö, provide a simple but realistic model for wave sizes above the mean sea level. Describe the model, sample space, and comment how the parameters could be estimated.*