NMST532 Design and Analysis of Medical Studies

On the Importance of Data Sampling Mechanisms

Michal Kulich

KPMS MFF UK

March 8, 2021

Parameter estimation

- Consider a collection of distributions $\mathcal{F}$, let $F_0 \in \mathcal{F}$ be the true distribution in the population we study.
- We want to estimate the parameter $\theta_0 = t(F_0)$ which lies in the parameter space $\Theta = \{t(F) : F \in \mathcal{F}\}$.
- Suppose we observe the data $\boldsymbol{X} = (X_1, \ldots, X_n)$ in order to estimate the parameter θ_0; define $\widehat{\theta}_n = \widehat{\theta}_n(\boldsymbol{X})$ an estimator of θ_0 based on the data $\boldsymbol{X}$.

Properties of parameter estimators

Statistical theory provides conditions that guarantee that the estimate we calculated from the data is in some sense *"correct"*:

1. $\widehat{\theta}_n$ is consistent, that is $\widehat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$ as $n \to \infty$;
2. $\widehat{\theta}_n$ has some asymptotic distribution, usually $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathrm{D}} \mathrm{N}(0, \Sigma)$ as $n \to \infty$;
3. a consistent estimator $S_n(\boldsymbol{X})$ of Σ is available.

If these conditions are valid we know that the estimator estimates the right thing and that we can correctly evaluate its uncertainty.

Properties of parameter estimators

- When the required conditions are violated, desirable properties of the estimator are lost.
- The worst consequences arise from losing the consistency of the estimator: then the estimator provides misleading results.
- Big Data is no remedy to lost consistency but the opposite. The more data are put into an inconsistent estimator the more strongly the analysis points to an incorrect solution.

Representative random sampling

- Suppose $X_1, \ldots, X_N$ is a random sample from the population of interest, that is, a collection of independent and identically distributed random objects with the distribution $F_0 \in \mathcal{F}$.
- Then we can obtain a consistent estimator of the parameter of interest $\theta_0 = t(F_0)$ quite easily, for example

$$\overline{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i$$

is a consistent estimator of $\theta_0 = \mathrm{E}\, X_i = \int x \, dF_0$.

Incomplete sampling

- However, in practice we may not get truly random samples that satisfy the above conditions.
- Suppose that the sample that we actually observe is $X_1, \ldots, X_n$, which is a subset of the representative random sample $X_1, \ldots, X_N$ of the size $n < N$.
- So, some observations from the representative sample are lost. Under what conditions we can use the subsample to get valid results about the parameter of interest θ_0?

Representation of incomplete sampling

- Divide the whole population of interest into strata (subgroups) according to the value of a discrete random variable $W \in \{1, \ldots, K\}$.
- Denote $\mathrm{P}[W = k] = \pi_k$.
- Denote the true conditional density of X given $W = k$ by $f_k(x)$.
- The true joint density of (X, W) at (x, k) is $f(x, k) = \pi_k f_k(x)$.
- The marginal density of X is $f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$ (a mixture of the conditional densities).
- Consider a random sample $(X_1, W_1), \ldots, (X_N, W_N)$ from the true joint density $f(x, k)$.

Representation of incomplete sampling

- Consider a random sample $(X_1, W_1), \ldots, (X_N, W_N)$ from the true joint density $f(x, k) = \pi_k f_k(x)$.
- Get a subsample from this ideal representative sample in such a way that the representation of the strata in the subsample is distorted:

We get to observe $(X_1, W_1), \ldots, (X_n, W_n)$ with the joint density $f^*(x, k) = q_k f_k(x)$.

- The marginal distribution of X in the subsample is

$$f^*(x) = \sum_{k=1}^{K} q_k f_k(x) \neq f(x) = \sum_{k=1}^{K} \pi_k f_k(x).$$

Consequences of incomplete sampling

- The marginal distribution of X in the subsample is

$$f^*(x) = \sum_{k=1}^{K} q_k f_k(x) \neq f(x) = \sum_{k=1}^{K} \pi_k f_k(x).$$

- In general, the distribution of X in the subsample is different from the distribution of X in the representative sample. So, we cannot consistently estimate any parameters of the desired distribution f. All estimators calculated on the subsample will be inconsistent.
- Under what circumstances is $f^*(x) = f(x)$? Need one of two conditions:
  - $q_k = \pi_k$, that is, the sample does not distort the distribution of W and is still representative for (X, W) or
  - $f_k(x) = f(x)$ for all k, that is, the distribution of X is the same in all strata, in other words, X is independent of W.

Incomplete sampling in regression

- Now consider a regression problem where the representative sample is $(Y_1, X_1), \ldots, (Y_N, X_N)$ from the joint density $f(y, x)$ and we are interested in parameters that describe the conditional density $f(y \mid x)$ — for example $\mathrm{E}\left[Y \,\middle|\, X\right]$.
- We have $f(y, x) = f(y \mid x) f(x)$, where $f(x)$ is the marginal density of the covariates.
- Add the strata and consider a random sample $(Y_1, X_1, W_1), \ldots, (Y_N, X_N, W_N)$ from the true joint density $f(y, x, k)$.
- Write the conditional density of (Y, X) given $W = k$ as

$$f_k(y, x) = f_k(y \mid x) f_k(x)$$

where $f_k(y \mid x)$ is the conditional density of Y given $X, W = k$ and $f_k(x)$ is the conditional density of X given $W = k$.

Incomplete sampling in regression

- The true conditional density of Y given X is

$$f(y \mid x) = \sum_{k=1}^{K} \pi_k f_k(y \mid x).$$

- The true joint density of Y and X is

$$f(y, x) = \sum_{k=1}^{K} \pi_k f_k(y \mid x) f_k(x).$$

- The true joint density of (Y, X, W) is

$$f(y, x, k) = \pi_k f_k(y \mid x) f_k(x).$$

Representation of incomplete sampling in regression

- Now consider an incomplete random sample $(Y_1, X_1, W_1), \ldots, (Y_n, X_n, W_n)$ from the density

$$f^*(y, x, k) = q_k f_k(y \mid x) f_k(x) \neq f(y, x, k) = \pi_k f_k(y \mid x) f_k(x).$$

- The marginal distribution of (Y, X) in the subsample is

$$f^*(y, x) = \sum_{k=1}^{K} q_k f_k(y \mid x) f_k(x) \neq f(y, x) = \sum_{k=1}^{K} \pi_k f_k(y \mid x) f_k(x).$$

- The conditional distribution of Y given X in the subsample is

$$f^*(y \mid x) = \sum_{k=1}^{K} q_k f_k(y \mid x) \neq f(y \mid x) = \sum_{k=1}^{K} \pi_k f_k(y \mid x).$$

Consequences of incomplete sampling in regression

- The conditional distribution of Y given X in the subsample is

$$f^*(y \mid x) = \sum_{k=1}^{K} q_k f_k(y \mid x) \neq f(y \mid x) = \sum_{k=1}^{K} \pi_k f_k(y \mid x).$$

- In general, the conditional distribution of Y given X in the subsample is different from the conditional distribution of Y given X in the representative sample. So, we cannot consistently estimate any regression parameters that describe the dependence of the response on the covariates.
- Under what circumstances is $f^*(y \mid x) = f(y \mid x)$? Need one of two conditions:
  - $q_k = \pi_k$, that is, the subsample does not distort the distribution of W and is still representative for (Y, X, W) or
  - $f_k(y \mid x) = f(y \mid x)$ for all k, that is, the conditional distribution of Y given X is the same in all strata, in other words, Y is conditionally independent of W given X.

The morale

- Consistent estimation is impossible unless the data are either complete or representative.
- Missing data (from any reason) compromise representativeness.
- Results obtained on data that was not collected by a thoughtfully developed sampling mechanism cannot be trusted.
- In regression problems, the data sampling mechanism may depend on the covariates but not on the response.

The morale (II)

- Data that “collects itself” (in various routinely maintained databases) cannot be trusted even for the simplest analysis tasks unless the database covers the whole population of interest.
- The belief that claims supported by “hard data” are automatically credible is inherently false.
- The argument that in the absence of representative data it is better to analyze an imperfect (meaning, *biased*) dataset to provide the desired results is thoroughly foolish and extremely dangerous.

The morale (III)

- The only way to obtain data that can provide reliable results is to collect the data actively according to some known sampling mechanism and to perform a careful verification of important data items. However, this principle is unpopular because it is time consuming, costly, and requires experts who have the brain to plan this. It seems the current world prefers quick solutions that can be done with little effort even if the resulting answers are worthless or directly harmful.
- Even when the data does not suffer from sampling problems, the associations identified in observational studies (without an active experimental component) cannot be interpreted as causal if the study was not randomized. This issue represents another layer of widespread misinterpretations of data analyzes.

Final remark

- The conclusions from this presentation will be portrayed as exaggerated. That cannot alter the fact that they are actually well justified and very rational.