MAT2013: Probability and Statistics [1] [2] **Spring 2017**

Lecture 6: Statistics: Introduction

Lecturer: Jae Guk, Kim *Scribes: Jisung Lim*

Scribes: Jisung Lim, *B.S. Candidate of Industrial Engineering in Yonsei University, South Korea.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Our study begins in this chapter with an introduction of statistics first as a conceptual framework complementary to—and dependent on—probability. We then provide a brief overview of the components of this "Statistical Analysis" framework, as a prelude to the detailed study of each of these components in the remaining chapters.

6.1 From Probability to Statistics

6.1.1 Random Phenomena and Finite Data Sets

Three entities for a systematic study of randomly varying phenomena:

1. X: *the actual variable of interest.*
   - This is the "Random Variable." It is an abstract, conceptual entity.
2. $\{x_i\}_{i=1}^{n}$: *n individual observations.*
   - A specific set out of many other possible *realizations of the random variable X.*
   - This set is commonly referred to as the "Data" and it is the only entity available in practice.
3. $f(x)$: *aggregate (or ensemble) description of the random variable.*
   - This is the theoretical model of 'how the probability of obtaining various results are distributed over the entire range of all possible values observable for X' and it is called "Probability Distribution Function (or PDF)."
   - We saw that it consists of a functional form, $f(x)$, and characteristic parameter, says $\boldsymbol{\theta}$; it is therefore more completely represented as $f(x|\boldsymbol{\theta})$, which literally reads "$f(x)$ *given* $\boldsymbol{\theta}$." (e.g., $\boldsymbol{\theta} = (\mu, \sigma^2)$ for Gaussian dist.)

What we've done:

- Probabilistic random phenomena analysis is based entirely on $f(x)$.
- This allowed us to abandon the impossible task of predicting an intrinsically unpredictable entity in favor of the more mathematically tractable task of computing the probabilities of observing the randomly varying outcomes.

- We, until now, have been assumed the availability of the complete $f(x)$, i.e., $f(x|\boldsymbol{\theta})$ with known $\boldsymbol{\theta}$.
- This allowed us to focus on the *first* task: computing probabilities and carrying out analysis, given any functional form and accompanying characteristic parameters, aggregately $f(x|\boldsymbol{\theta})$, assumed known.

Our natural question is:

- Where either the functional form $f(x)$, or the specific characteristic parameter values $\boldsymbol{\theta}$ come from.
- In actual practice, what is really available about any random variable of interest?
- How does one go about obtaining the complete $f(x)$ required for random phenomena analysis?

Actually, in practice:

- For any specific random varying phenomenon of interest, the theoretical description, $f(x)$, is never completely available.
- This is usually because the characteristc parameters assocated with the particular random variable in question are unknown.
- In practice, only finite data in the form of a set of observations $\{x_i\}_{i=1}^n$ is available.

The problem at hand is:

- To apply the theory of random phenomena analysis successfully to practical problems, how the complete $f(x)$ to be determiend from finite data?
- How to analyze randomly varying phenomena on the basis of finite data sets?

Now, this is the domain of Statistics! Then, what is statistics?

- In a sense of the reverse problem of probability.
  - With probability, $f(x)$—the functional form along with the parametes —is given, and analysis involves determining the probabilities of obtaining specific observations $\{x_i\}_{i=1}^n$ from the random variable X.
  - The reverse is the case with statistics: $\{x_i\}_{i=1}^n$ is given, and analysis involves determining the appropriate (and complete) $f(x)$—the functional form and the parameters—for the random variable X that generated the given specific observation.
- The theoretical concept of the pdf, $f(x)$, plays a significant role in determining, from the finite data set, the most likely underlying complete $f(x)$, and to qunatify the associated uncertainty.
- In this sense, "Statistics" is referred to as a methodology for
  - *inferring* the characteristics of the complete pdf, $f(x)$, from finite data set $\{x_i\}_{i=1}^n$,
  - and *qunatifying* the associated uncertainty.
- Our focus of interest is
  - In *Probability*, the "Random Variable", says X, as an abstract entity.
  - In *Statistics*, the finite "Data Set", says $\{x_i\}_{i=1}^n$, as a specific realization of X.

Example 6.1.1. Three times coin tossing

What we've been doing before:

1. Establishing probability space
   - Random experiment: Tossing a coin three times.
   - Sample space: $\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$
   - *σ-algebra*: $\mathcal{F} = 2^{\Omega}$
   - Probability measure: $\mathbb{P}[A] = \frac{|A|}{|\Omega|} \quad \forall A \in \mathcal{F}$ ($|\cdot|$: cardinality of a set).
2. Define random variable $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$

$$X = \text{The number of observed tails in 3 tossing}$$

3. Define P_X

$$\forall B \in \mathcal{B}, \quad P_X(B) = \mathbb{P}[X^{-1}(B)] = \mathbb{P}[\{\omega \in \Omega : X(\omega) \in B\}]$$

4. But one may skip the steps 1–3 if a random phenomenon has pre-existing "canned" model. In this case, this specific randomly varying phenomenon follows the binomial distribution with $n = 3$ and p as the characteristic parameter. The ensemble description is the binomial pdf:

$$f(x|p) = \binom{3}{x} p^x (1-p)^{3-x}; \ x = 0, 1, 2, 3$$

   Eventually, we've found the theoretical ensemble description $f(x)$.

5. Let us set the parameter p with an ideal value $\frac{1}{2}$, then

$$\begin{aligned} P(X = 0) &= 1/8 \\ P(X = 1) &= 3/8 \\ P(X = 2) &= 3/8 \\ P(X = 3) &= 1/8 \end{aligned}$$

In practice:

1. Observe the data:
   Let us observe the result S of 10 independent random experiments as a specific experimental realization of the random variable X.

$$S = \{0, 1, 3, 2, 2, 1, 0, 1, 2, 2\}$$

2. Evaluate relative frequency:

Strictly from the limited data, the various relative frequencies of occurrence are given by

$$\begin{aligned} F_{\text{rel}}(0) &= 0.2 \\ F_{\text{rel}}(1) &= 0.3 \\ F_{\text{rel}}(2) &= 0.4 \\ F_{\text{rel}}(3) &= 0.1 \end{aligned}$$

3. Approximate the true probability distribution P_X with F_{rel}:
   If one can assume that this data set is "representative" of typical behavior of the random variable, the observed relative frequency F_{rel} can be considered as an approximate representation of true probability distribution P_X.

Hence, the implication is that: the data set appears to be somewhat **consistent with the theoretical model** when $p = 0.5$. $\Diamond$

6.1.2 Finite Data Sets and Statistical Analysis

Three concepts are central to statistical analysis: population, sample, and statistical inference.

1. *Population*: This is the complete collection of all the data obtainable from a random variable of interest.
   - Clearly, it is impossible to 'realize' the population in actual practice.
   - But as a conceptual entity, it serves a critical purpose: *The full observational "realization"* of the random variable X.
   - It is to statistics what the sample space is to probability theory.
2. *Sample*: A *specific* set of actual observations obtained upon performance of an experiment.
   - By definition, this is a finite subset of data selected from the population of interest.
   - It is the only information that is actually available *in practice.*
3. *Statistical Inference*: Any statement made about the population on the basis of information extracted from a sample.
   - Because the sample will never encompass the entire population, such statements must include a measure of the "unavoidable associated uncertainty".

In the probability theory

- In the probability theroy, we focus on the random variable X.
- Then we utilizes the sample space $(\Omega, \mathcal{F})$ and the state space $(\mathbb{R}, \mathcal{B})$ as the basis for developing the theoretical ensemble description $f(x)$.

In practice

- For any specific problem, the focus shifts to the actual observed data, $\{x_i\}_{i=1}^n$.
- The equivalent conceptual entity in statistics becoems the population—the observational ensemble to be described and characterized.

- While the sample space of the probability theory can be specified *a-priori*, but the population in statistics refers to observational data, making it a specific, *a-posteriori* entity.

Example 6.1.2. 3 coin tossing revisit

What we can know before any experiment.

Before any actual experiments are performed, we know the support of a random variable, given by

$$\textit{support}: V_X = \{x \in \mathbb{R} : f_X(x) > 0\} = \{0, 1, 2, 3\}.$$

- V_X indicates that with each performance of the experiment, the outcome will be one of the 4 numbers contained in it.
- We can compute probabilities of observing any one of the 4 possible alternatives.

The unknown characteristic parameter

For the generic coin with parameter p, we are able to obtain an explicit expression for how the outcome probabilities are distributed over the values in V_X.

- For a specific coin for which, say, $p = 0.5$, we can compute the probabilities of obtaining 0, 1, 2 or 3 tails, as we just did in *Example* 6.1.1.
- But in practice, the true value of p associated with a specific coin is not known *a-priori*; it must be determined from experimental data.

Finite sample from infinite population

We may consider a single, 10-observation sample for the specific "coin in question", given by

$$S_1 = \{0, 1, 3, 2, 2, 1, 0, 1, 2, 2\}$$

.

- A specific sample is considered to be drawn from the conceptual population of all such data obtainable from this specific coin characterized by the true, but unknown, value p.
- Although finite, S_1 contains information about the true value of the characteristic parameter p associated with this particular coin.
- Determining appropriate estimates of this true value is a major objective of statistical analysis.

Different sample from the same population

Because of the finiteness of sample data, a second series of such experminets will yield a *different* set of results, for example,

$$S_2 = \{2, 1, 1, 0, 3, 2, 2, 1, 2, 1\}$$

- This is another sample from the same population, but as a result of inherent variability, it is different from S_1.
- Nevertheless, this new sample S_2 also contains information about the unknown characteristic parameter, p.

How about another population

Next consider *another* coin, says that being characterized by $p = 0.8$, and suppose that after performing the 3 coin toss experiment, say $n = 12$ times, we obtain:

$$S_3 = \{3, 2, 2, 3, 1, 3, 2, 3, 3, 3, 2, 2\}$$

- This set of results is considered to be just one of an infinite number of other samples that could potentially be drawn from the population characterized by $p = 0.8$.
- As before, it also contains information about the value of the unknown population parameter.

Summary

1. *Invariable sample space*:
   - With probability, the support of the random variable for this example is finite, specifiable *a-priori.*
   - It remains as given no matter what the value of the population parameter p is.
2. *Variable population*:
   - Not so with the population. It is finite and its elements depend on the specific value of the characteristic population parameter.
   - In the case of $p = 0.8$, the relative rarity of '1' in the sample S_3 indicates that the population of all possible observations from the 3 coin toss experiment will very rarely contain the number '1'.
   - Information about the true, but unkown, value of the characteristic parameter, p, associated with each coin's population is contained in each finite data set.

$\Diamond$

In a general sense, the cases that we will confront til now are

1. The form of the PDF $f(x)$ is known, but the parameter $\boldsymbol{\theta}$ is unknown.
2. Data is available, but in the form of only finite-sized samples.
3. Whereas, the full ensemble characterization we seek is of the entire population.

That is, we are left with no other choice but to use the samples, even though finite in size, to characterize the population.

- **The population**—the "full observational manifestation" of the random variable X—is that ideal, conceptual entity one wishes to characterize.
- The objective of random phenomena analysis is to **characterize the population** completely with the pdf $f(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents the characteristic parameters of the specific population in question.
- However, because it is impossible to realize the population in its entirety via experimentation, one must therefore settle for characterizing it by drawing **statistical inference** from a *finite sample subset.*
- Clearly, the success of such an endeavor depends on *the sample* being **"representative" of the population**.

Statistics therefore involves not only systematic *analysis* of data, but also systematic data *collection* in such a way that the sample truly reflects the population, thereby ensuring that the sought-after information will be contained in the sample.

References

[1] Youngstown State University G. Jay Kerns. *Introduction to Probability and Statistics Using R.* `http://ipsur.org/index.html`. G. Jay Kerns, 2010.

[2] Babatunde A. Ogunnaike. *Random Phenomena: Fundamentals of Probability and Statistics for Engineers.* CRC Press, 2009.