Statistical Inference

January 6, 2019 9:08 PM

Basics: Introduction

- Statistical inference is not so much about the methods of statistics but the "why".
- what is statistics as a subject all about?
- statistical methods are used in:
  - Finance
  - Machine learning
  - medicine
  - quantum physics

    ⋮

    more!
- Furthermore, "statistical reasoning" is becoming more and more important!
- It is being used as a tool to reason about reality.
- Note: significant decisions are made based on statistical analysis.
- So we want the rules of statistical reasoning to be ____ = logical, free of contradictions, ____, etc ... so we feel confident that whatever the conclusion / inference we draw makes sense.
- Current state of statistics
  - Many different points of view about what the the correct statistical reasoning is.
  - This makes learning the subject hard.
- Purpose of this course (STAC58 - Statistical Inference)

  1.) Survey the various approaches

  2) present the outline of a logical way to develop a theory of statistical reasoning.
- Some phenomenon / context in the real world that we have questions about
- Questions like: 1) what is the value of some quantity of interests?

  eg. mean half life length of a neutron

  2) Does a certain quantity take a particular value?
- when can statistical inference play a role?

Statistical Problems

- The first thing we need to do is be very clear about what a statistical problem is.
- It is all based on "measuring" and counting.
- we have a population $\underline{\Omega}$ = a finite set of objects of interests.

Eg. $\underline{\Omega}$ = set of all students enrolled at UofT on Jan 2, 2015

- $\underbrace{\#}_{}(\Omega) < \infty$

cardinality / # of items in the set

- we have a measurement(s) defined on Ω

$X : \Omega \to \mathcal{X}$

- for $\omega \in \Omega$ = set of students at UofT.

Define

$X_1(\omega)$ = height of ω in cm (interval)

$X_2(\omega)$ = weight of ω in kg (interval)

$X_3(\omega)$ = gender of ω (categorical)

- $X = (X_1, X_2, X_3) : \Omega \to R \times R \times \{M, F\}$

- Ω and X define relative frequency function over x.

$$f_X(x) = \frac{\#\{\omega : X(\omega) = x\}}{\#(\Omega)}$$

= proportion of individuals in Ω whose x measurements is $x \in \mathcal{X}$.

- note: i) $0 \le f_X(x) \le 1$

ii) $\sum_{x \in \mathcal{X}} f_X(x) = 1$

and only finitely many $x \in \mathcal{X}$ have $f_X(x) > 0$.

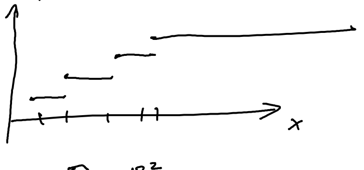
- when $\mathcal{X} = R$ (or an interval)

$$F_X(x) = \frac{\#\{\omega : X(\omega) \le x\}}{\#(\Omega)}$$ = cumulative distribution function of X (CDF of X)

$$= \sum_{z \le x} f_X(z)$$

$f_X(x) = F_X(x) - F_X(x-0)$, where $F_X(x-0) = \lim_{z \uparrow x} F_X(z)$

- So F_X and f_X are two equivalent ways of presenting a frequency distribution.

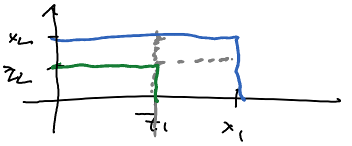


- when $\mathcal{X} = \mathbb{R}^2$

$$F_X(x_1, x_2) = \frac{\#\{\omega : X_1(\omega) \le x_1, X_2(\omega) \le x_2\}}{\#(\Omega)}$$

$$= \sum_{\substack{z_1 \le x_1 \\ z_2 \le x_2}} f_X(z_1, z_2)$$

$$f_X(x_1,x_2) = \lim_{z_1 \uparrow x_1} \left[F_X(x_1,x_2) - F_X(x_1, z_2) - F_X(z_1,x_2) + F_X(z_1,z_2) \right]$$



So, $F_X \longleftrightarrow f_X$

- The whole point of any statistical analysis is to learn something about F_X.
- how do we do this?
- If possible we do a census, namely compute $X(\omega)\ \forall \omega \in \Omega$ of the form f_X.
- Typically count (return to this in a moment)
- why do we want to know F_X?

eg relationships among variables.

- Suppose (X, Y), where $X: \Omega \to X$, $Y: \Omega \to Y$ and we want to know if there is a relationship between X & Y on Ω.
- form the conditional relative frequency distribution.

$$f_{Y|X}(y|x) = \frac{\#\{\omega \mid X(\omega) = x,\ Y_2(\omega) = y\}}{\#\{\omega \mid X(\omega) = x\}}$$

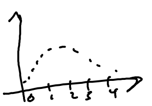
$$= \frac{f_{(X,Y)}(x, y_2)}{f_X(z)}$$

Definition: X and Y are related variables over Ω if $f_{Y|X}(\cdot|x)$ changes as x changes.

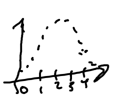
- The "form" of the relationship between X and Y is given by how $f_{Y|X}(\cdot|z)$ changes as x_2 changes.

eg. Ω = 1st year students at U of T
Y = GPA as of Dec 31, 2015.
X = gender

$f_{Y|X}(\cdot|m)$



$f_{Y|X}(\cdot|F)$



- often simplifying assumptions are introduced.
- regression assumption: $f_{Y|X}(\cdot|x)$ changes at most through its mean as x changes, $E(Y|X)(x)$

$$= \frac{1}{\#\{\omega : X(\omega) = x\}} \sum_{\{\omega : X(\omega) = x\}} Y(\omega)$$

$$\overset{Ex}{=} \sum_y y\, f_{Y|X}(y|x)$$