Statistical Inference

Real Analysis in E+R

(I) Basics

(1.) Introduction

- not so much about the methods of statistics but the why
- what is statistics as a subject all about?
- statical methods are used
  - finance
  - machine learning
  - medicine
  - quantum physics.
  - ⋮
- "statistical reasoning" is becoming more important
- it is being used as a tool to reason about reality
- significant decisions are made based on statistical analysis
- so we want the rules of statistical reasoning to be sound = logical, free of contradictions, paradoxes, etc. so we feel confident that whatever conclusions / inferences

that we draw make sense

- current state of statistics
  - many different points of view about what is correct statistical reasoning
  - makes learning the subject hard.
- purpose of the course
  (1) survey the various approaches
  (2) present the outline of a logical way to develop a theory of statistical reasoning
- start from the start
- some phenomenon/context in the real world that we have questions about
- questions like (1) what is the value of some quantity of interest? eg. mean half life length of a neutron

(2) does a certain quantity take a particular value?

- when can statistical inference play a role?

② 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 Ω = a finite set of objects of interest

eg Ω = set of all students enrolled at UofT on Jan. 2, 2016

- $\#(\Omega) < \infty$.
- we have a measurement defined on Ω

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

eg (cont'd) - for $\omega \in \Omega$ = a student at UofT

- define $X_1(\omega)$ = ht of ω in cm (interval)

$X_2(\omega)$ = wt of ω in kg (")

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

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

- Ω and X define a **relative frequency function** over $\mathcal{X}$

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

= proportion of individuals in Ω whose X measurement 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 $\mathcal{X} = \mathbb{R}$ (or an interval)

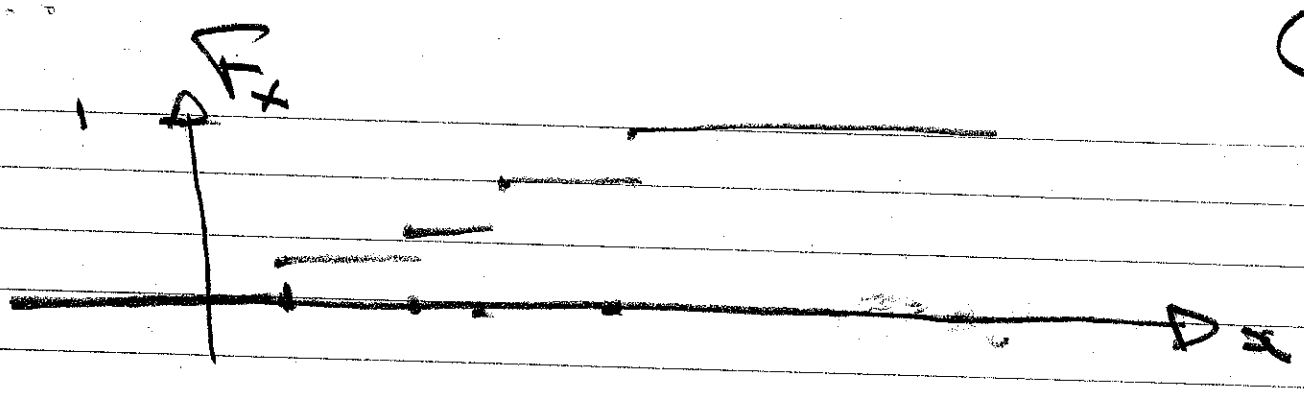
$$F_X(x) = \frac{\#\{\omega : X(\omega) \le x\}}{\#(\Omega)} = \text{cumulative distribution function}$$

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

$$f_X(x) = F_X(x) - F_X(x-0)$$

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.

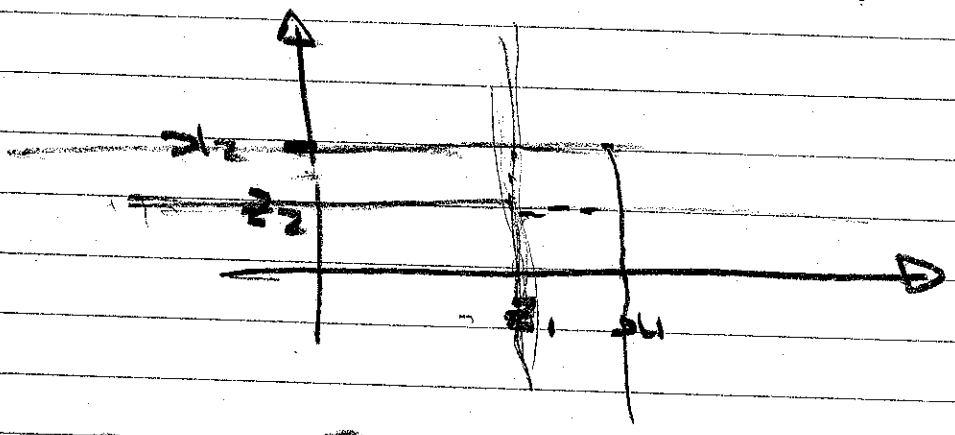


- idea $\Omega = \mathbb{R}^2$

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

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

$$f_X(x_1, x_2) = \lim_{\substack{z_1 \to x_1 \\ z_2 \to x_2}} \left(F_X(x_1, x_2) - F_X(x_1, z_2) - F_X(z_1, x_2) + F_X(z_1, z_2) \right)$$



so $F_X \Rightarrow f_X$

- the whole point of any statistical analysis is to learn something about R_X

- how do we do this?

- if possible we could do a census, namely compute $X(\omega)$ $\forall \omega \in \Omega$ and then form F_X

- typically can't (return to this in a moment)

- why do we want to know F_X?

relationships among variables

- suppose (X, Y), where $X: \Omega \to \mathbb{R}_X$, $Y: \Omega \to \mathbb{R}_Y$, and we want to know if there is any relationship between X and Y on Ω

- form the empirical relative frequency distribution

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

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

what makes Y a response and X a predictor?

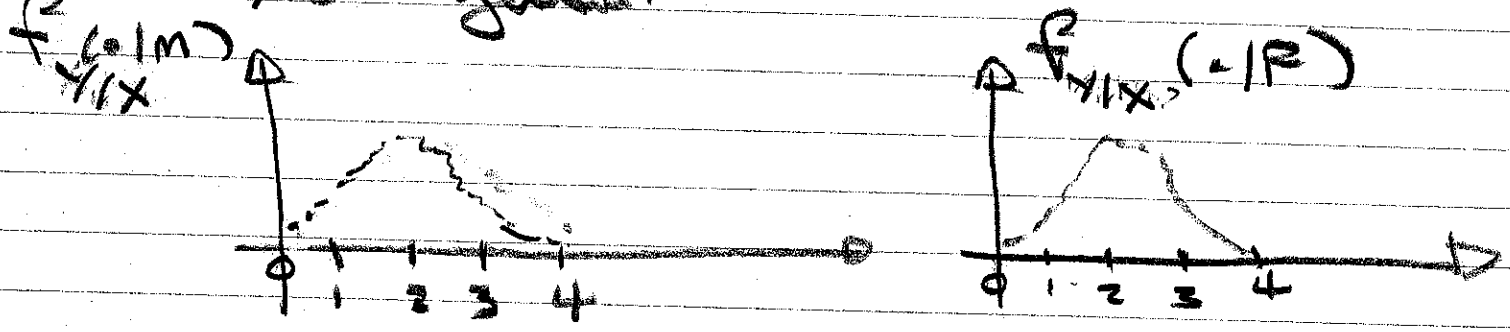
Def X and Y are related variables over Ω if $f_{Y|X}(\cdot|x)$ changes as x changes

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

eg Ω = 1st yr students at UofT

Y = GPA as of Dec 31, 2015

X = gender



- often simplifying assumptions are introduced

- regression assumption $f_{Y|X}(\cdot|x)$ changes at most through its mean as x changes, $E(Y|X=x)$

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

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

- linear regression assumption $E(Y|x)$ is a linear function of x

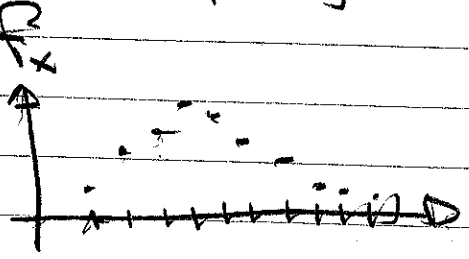
③ Infinity and Continuity

- all populations are finite

- so all relative frequency distributions are positive on a finite number of points

⇒ Ω = subsets of UofT

$X(\omega)$ = ht in cm



- sometimes a measurement can be thought of as being taken to an ever finer accuracy

- histogram - divide up range into k intervals

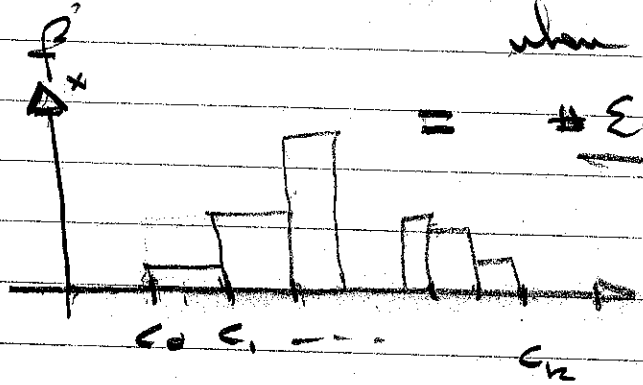
$c_0 < c_1 < \dots < c_k$

- density histogram

$$f_X^{\text{dens}}(x) = \sum_{x \in (c_i, c_{i+1})} f_X(x \in \omega) / (c_{i+1} - c_i)$$

when $x \in (c_i, c_{i+1}]$

$$= \frac{\#\{\omega : c_i < X(\omega) \le c_{i+1}\}}{\#(\Omega)(c_{i+1} - c_i)}$$



- then for $c_j > c_i$

$$F_X(c_j) - F_X(c_i) = \int_{c_i}^{c_j} f_X^{\text{dens}}(x)\,dx$$

- for such measurements we can introduce the idea of a continuous approximation

- let f be a prob. density fn s.t.
  (i) $f(x) \geq 0$ (ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

- then f approximates f_X^{dens} if $\forall a < b$
  $$\int_a^b f(x)\,dx \approx \int_a^b f_X^{\text{dens}}(x)\,dx.$$

- so continuous distributions (on infinite sample space) arise as approximations

  make precise via limits as measurements are made to finer and finer accuracy

Ex Ω = students at UofT

- $X(\omega)$ = ht in cm.

- plausibly we consider the approximation
  $X(\omega) \sim N(\mu, \sigma^2)$
  for some $\mu \in \mathbb{R}$, $\sigma^2 \in (0, \infty)$

- actually $\mu = \frac{1}{\#(\Omega)} \sum_{\omega \in \Omega} X(\omega)$
  $\sigma^2 = \frac{1}{\#(\Omega)} \sum_{\omega \in \Omega} (X(\omega) - \mu)^2$

eg Ω = students at UofT

- $Y(\omega)$ = wt in gms

  $X(\omega)$ = ht in cms

- $\begin{pmatrix} X(\omega) \\ Y(\omega) \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}\right)$

$$\mu_X = \frac{1}{\#(\Omega)} \sum_{\omega\in\Omega} X(\omega), \quad \mu_Y = \frac{1}{\#(\Omega)} \sum_{\omega\in\Omega} Y(\omega)$$

$$\sigma_X^2 = \frac{1}{\#(\Omega)} \sum_{\omega\in\Omega} (X(\omega) - \mu_X)^2$$

$$\rho_{XY} = \frac{1}{\sigma_X\sigma_Y} \frac{1}{\#(\Omega)} \sum_{\omega\in\Omega} (X(\omega)-\mu_X)(Y(\omega)-\mu_Y)$$

$$= \frac{\sum_{\omega\in\Omega} (X(\omega)-\mu_X)(Y(\omega)-\mu_Y)}{\sqrt{\sum_{\omega\in\Omega} (X(\omega)-\mu_X)^2}\sqrt{\sum_{\omega\in\Omega} (Y(\omega)-\mu_Y)^2}}$$

doesn't matter $\#(\Omega)$

- $Y(\omega) \mid X(\omega) = x \sim N\left(\mu_Y + \sigma_Y \rho_{XY} \frac{(x-\mu_X)}{\sigma_X}, \sigma_Y^2(1-\rho_{XY}^2)\right)$

- so if $\rho_{XY} \neq 0$ relationship between X and Y exists

- how strong is the relationship?

- the closer ρ_{XY}^2 is to 1 the stronger the relationship.

- note we can write

$$E(X_1 \mid X_2 = x_2)$$

$$= \mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_2 - \mu_2)$$

$$= \left(\mu_1 - \frac{\sigma_1}{\sigma_2}\rho\mu_2\right) + \frac{\sigma_1}{\sigma_2}\rho x_2$$

$$= \beta_0 + \beta_1 x_2$$

where $\beta_0 = \mu_1 - \frac{\sigma_1}{\sigma_2}\rho\mu_2$, $\beta_1 = \frac{\sigma_1}{\sigma_2}\rho$

so alternatively we can write

$$Y = \beta_0 + \beta_1 x + \varepsilon \quad \text{where } \varepsilon \sim N(0, \sigma^2)$$

- note - if we know f_X we know everything from a statistical point-of-view

- note - what is a relationship between X and Y exactly

- must be able to assign any value of X to any $\omega \in \Omega$.

(3) Sampling

- typically we can't collect a census
- so select $n < \#(\Omega)$ and $\{\omega_1, \ldots, \omega_n\} \subseteq \Omega$ obtaining $x_1 = X(\omega_1), \ldots, x_n = X(\omega_n)$
- based on the data $x = (x_1, \ldots, x_n)$ we make inferences about F_X
- how do we select $\{\omega_1, \ldots, \omega_n\}$?
- clearly we want to avoid selection effects as this will "bias" our results

Eg - if $\omega_i \in \Omega$ = UofT students, x always female then x would not be representative

- solution - use a random mechanism to select $\{\omega_1, \ldots, \omega_n\}$

Eg - put $\#(\Omega)$ chips in a bowl each labelled with a number in $\{1, 2, \ldots, \#(\Omega)\}$ where i corresponds to i-th member of Ω

- stir up the chips and pick one without looking say the one labelled i
- continue without replacement

sampling with and without replacement

- the select population count:

- we call this via probability

$$P(X(\omega_1) = x_1) = \frac{\#\{\omega : X(\omega) = x_1\}}{\#(\Omega)}$$

$$= f_X(x_1)$$

$$P(X(\omega_2) = x_2 \mid X(\omega_1) = x_1)$$

$$= \begin{cases} \dfrac{\#\{\omega : X(\omega) = x_2\} - 1}{\#(\Omega) - 1} & x_2 = x_1 \\ \dfrac{\#\{\omega : X(\omega) = x_2\}}{\#(\Omega) - 1} & x_2 \neq x_1 \end{cases}$$

$$= \begin{cases} \dfrac{f_X(x_2) - 1/\#(\Omega)}{1 - 1/\#(\Omega)} & x_2 = x_1 \\ \dfrac{f_X(x_2)}{1 - 1/\#(\Omega)} & x_2 \neq x_1 \end{cases}$$

$$\approx f_X(x_2)$$

$$\therefore P(X(\omega_1) = x_1, X(\omega_2) = x_2) = P(X(\omega_1) = x_1) P(X(\omega_2) = x_2 \mid X(\omega_1) = x_1)$$

$\propto f_X(x_1) f_{X_2}(x_2)$

- provided $n \ll \#(\Omega)$

$P(X(\omega_1) = x_1, \ldots, X(\omega_n) = x_n)$

$= f_X(x_1) \cdots f_X(x_n)$ approximately

- i.e. $x_1, \ldots, x_n$ are i.i.d. f_X

- note: if we know f_X we know everything from a statistical point-of-view.

- note - if we have generated the data via a random mechanism then we are justified in referring to the data as being objective

(4) Statistical Models

- for inference about f_X we have the data $x = (x_1, \ldots, x_n)$
- some basic inferences justified by convergence results

e.g. $A \subseteq \mathcal{X}$

- $P(A) = \sum_{x \in A} f_X(x) =$ proportion of $\omega \in \Omega$ s.t. $X(\omega) \in A$
- estimate by $\hat{P}(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(x_i)$

$=$ proportion of sample values s.t. $x_i \in A$

$\xrightarrow{\text{SLLN}} P(A)$ as $n \to \infty$

- $I_A(X(\omega_i)) \sim \text{Bernoulli}(P(A))$ when ω randomly selected
- $\text{Var}(\hat{P}(A)) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(I_A(X(\omega_i)))$

$= \frac{1}{n} \text{Var}(I(X(\omega))) = \frac{P(A)(1-P(A))}{n}$

- actually $n\hat{P}(A) \sim \text{Binomial}(n, P(A))$
- $\frac{\sqrt{n}(\hat{P}(A) - P(A))}{\sqrt{\hat{P}(A)(1-\hat{P}(A))}} \xrightarrow{\text{CLT}} N(0,1)$
- so $\hat{P}(A) \pm 3\sqrt{\frac{\hat{P}(A)(1-\hat{P}(A))}{n}}$ contains $P(A)$ with visual certainty

- but requires n "large"
- similarly for other characteristics of f_X e.g. μ_X, σ_X^2, σ_X/μ_X, etc.
- want inference methods that don't require large n.
- typically this involves making assumptions
- assumptions involve making (subjective) choices

Principle of Empirical Criticism

Any ingredient (assumption) we use as part of a statistical analysis must be checkable against the (objective) data to ensure that it makes sense.

- the most important assumption made is the choice of a statistical model
- we assume $f_X \in \{f_\theta : \theta \in \Theta\} = \mathcal{M}$ where f_θ is a density on $\mathcal{X}$ for each $\theta \in \Theta$
- θ is called the parameter of the model and Θ is called the parameter space
- typically θ indexes; to each value of

- θ then corresponds a unique density f_θ (no nonidentifiability)

- we have to check the assumption

$$f_x \in \{f_\theta : \theta \in \Theta\} - \Pi$$

- how? later

- so suppose we have checked Π against x and have decided Π is okay (note: in general we can never say it is correct)

- then we want rules to apply to (Π, x) to make inferences about true value of θ (or the equivalently the true f_x)

eg Ω = students at UofT

- $X(\omega)$ = ht of ω in cm.

- assume $X(\omega) \sim N(\mu, \sigma^2)$

- $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$

- to check the model look at standardized residuals,

- $r_i = \frac{x_i - \tilde{x}_i}{\|x_i - \tilde{x}_i\|} \overset{iid}{\sim} \text{Uniform}(S^{d-1}(1) \cap S^n)$

- propose test statistics that look for patterns.

⇒ theorem $T(r) = \frac{1}{n}\sum_{i=1}^{n} r_i^3 \xrightarrow{d} 0$

so should be close to 0

- why make this assumption $F_X \in M$

(1) small n so assumption replaces data for more accurate inference

(2) enables the development of a theory of inference as opposed to assumptions

5 Types of Inference

- often we don't need to know θ but only $\eta = \eta(\theta) \in \mathbb{R}$

Eg - $x = (x_1, \ldots, x_n) \sim N(\mu, \sigma^2)$

$(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$

- $\eta(\mu, \sigma^2) = \mu$ (inference about mean)

$\eta(\mu, \sigma^2) = \sigma$ (inference about standard deviation)

$\eta(\mu, \sigma^2) = \sigma/\mu$

$\eta(\mu, \sigma^2) = \mu + \sigma z_{.75}$ (3rd quartile)

$\eta(\mu, \sigma^2) = \Phi\left(\frac{x_2 - \mu}{\sigma}\right) - \Phi\left(\frac{x_1 - \mu}{\sigma}\right)$

$=$ prob cont of (x_1, x_2)

- Two types of inference (this is what we'll want a theory of inference to give us)

(1) estimate $\hat{\eta}(x)$ together with an assessment of the accuracy of the estimate as given by a set $C(x) \subseteq \mathbb{R}$ with $\hat{\eta}(x) \in C(x)$ and a size of $C(x)$ gives assessment

(2) assess hypothesis $H_0 : \theta \in \Theta_0 \subseteq \Theta$ (i.e. we have a hypothesis about the true value of θ is $\theta_0 \in \Theta$) by quoting a measure of the evidence that H_0 is true together with a measure of the strength of this evidence.

- all theories of statistical inference attempt to answer (1) and (2)

(6.) Statistical Analyses

- we can conceive of a statistical analysis as following these steps

identify the population and the variables being measured

↓

(subjective) choose a model M and any other ingredients required for application of the theory

↓

note

(objective) generate data x by random sampling from the relevant population.

↓

check the ingredients you have chosen against x

okay → apply theory of inference

→ modify ingredients