STAC58 - Chapter 1

Textbook: Probability and Statistics - Evans & Rosenthal,

Chapter 5.9

Measuring Statistical Evidence using relative belief

- M. Evans

website: http://www.utstat.utoronto.ca/mikevans/stac58/stac58.html

Evaluations - Midterm 40%
- final 60%.

January 7, 2019
11:15 AM

STAC58 - Statistical Inference

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 **sound** = logical, free of contradictions, **paradoxes**, 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 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 interest?
  eg. mean half life length of a neutron
  Answer: An estimate of assessment of its error
  2) Does a certain quantity take a particular value?
  Answer: hypothesis assessment - evidence for or against and a measure of strength.
- when can statistical inference play a role?

Theory tells how accurate estimate is.

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 interests.

Eg. Ω = set of all students enrolled at UofT on Jan 7, 2019

- $\#(\Omega) < \infty$

cardinality / # of items in the set

- we have a measurement(s) defined on Ω

$X : \Omega \to \mathcal{X}$ $\quad \omega \in \Omega \; X(\omega)$

- 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)

$x(\omega) = \begin{pmatrix} X_1(\omega) \\ X_2(\omega) \\ X_3(\omega) \end{pmatrix}$

$x(\omega)$ is 3d measurements

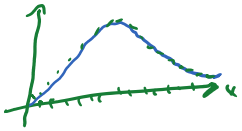
- $X = (X_1, X_2, X_3) : \Omega \to R \times R \times \{M, F\}$

$X : \Omega \to R' \times R' \times \{M, F\}$ ($\mathcal{X}$)

- Ω and X define relative frequency distribution over X.

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

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



simplify by introducing continuous approximation

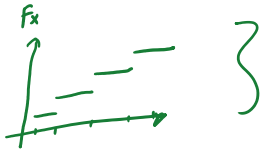
discrete is too hard from! approx it.

- note: i) $0 \le f_X(x) \le 1 \quad \forall x \in \mathcal{X}$
  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)

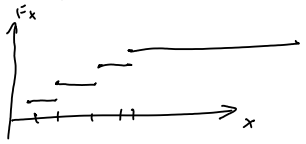
$= \sum_{z \le x} f_X(z)$



Step function

$F_X(x) = F_X(x) - F_X(x - \underbrace{\epsilon}_{\text{epsilon}})$, where $F_X(x-) = \lim_{\epsilon \to 0} F_X(\epsilon)$

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

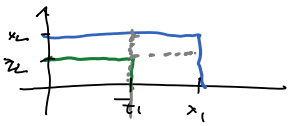


- when $\mathcal{X} = \mathbb{R}^2$

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

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

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



So, $F_X \leftrightarrow f_X$

- The whole point of any statistical analysis is to learn something about F_X.
- how do we do this?
- If possible we do a <u>census</u>, 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 $X: \Omega \to \mathcal{X}$, $Y: \Omega \to \mathcal{Y}$ and we want to know if there is a relationship between X & Y on Ω.
- form the conditional relative frequency distribution.

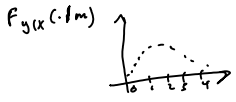
★ $f_{Y|X}(y|x) = \dfrac{\#\{\omega | X(\omega) = x,\ Y_2(\omega) = y\}}{\#\{\omega : X(\omega) = x\}}$

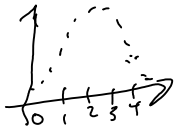
$$= \frac{f_{(X,Y)}(x, y_2)}{f_X(x)}$$

★ <u>Definition</u>: 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 $f_{Y|X}(\cdot|x_2)$ changes as x_2 changes.

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





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

$$\xrightarrow{\text{LLN}} \sum_y y\, f_{Y|X}(y|x)$$

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

$f_{y|x}(\cdot|F)$

eg. relationship among variables

- Suppose $(x, y) : \Omega \to \mathcal{X} \times \mathcal{Y}$ and we want to know if there is any relationship between X & y.
- Form the conditional relative frequency distribution

★ $$F_{y|x}(y|x) = \frac{\#\{\omega : x(\omega) = x,\ y(\omega) = y\}}{\#\{\omega : x(\omega) = x\}} = \frac{F_{(x,y)}(x,y)}{f_x(x)}$$

★ **Def** X & Y are **related variables** over a population Ω, if $F_{y|x}(\cdot|x)$ changes as X changes.

eg. Ω = students at uof T

$x(\omega)$ = gender $\quad y(\omega)$ = GPA

$f_{y|x}(\cdot|M)$

$f_{y|x}(\cdot|F)$

- The form of the relationship

★ if x and y have no relationship then

$f_{y|x}(y|x) = f_y(y)$

$\iff F_{(x,y)}(x|y) = F_{x(x)} f_y(y)$

- the form of the relationship between X & y when it exists is given by how $F_{y|x}(\cdot|x)$ changes with x

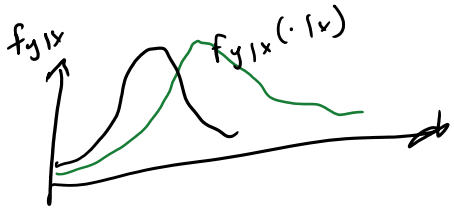
- Often simplifying assumptions are made

- regression assumptions

  - here $y = \mathbb{R}$ and we assume

    $f_{y|x}(\cdot|x)$ changes at most through

  ★ $E(y|x)(x) = \frac{\sum_{\omega : x(\omega)=x} }{\#\{\omega : x(\omega)=x\}}$ = average value of y in the sub population $\{\omega : x(\omega) = x\}$



- regression assumption

  - same distribution but shifted.

Linear regression assumption

$E(y|x) \in [\{g_1, \dots, g_n\}$ where

$g_i : x \to \mathbb{R}$ $i = 1, \dots, n$.

i.e $E(y|x)(x) = \beta_1 g_1(x) + \dots + \beta_n g_n(x)$, for some $\beta_1, \dots, \beta_n \in \mathbb{R}$

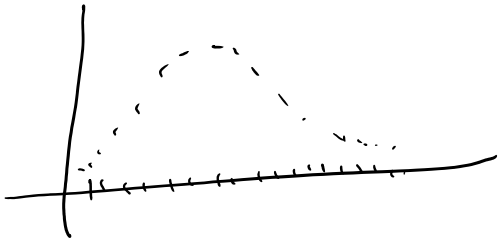
eg $g_1(x) = 1$, $g_2(x) = x$, $g_3(x) = x^2$

Infinity and Continuity

- Recall, all populations are finite
- As a result, all relative frequency distribution are positive on a finite # of points

eg. Ω = students at the university of Toronto

$x(\omega)$ = height in centimeters.



- Sometimes measurements can be thought of as being taken to an ever finer accuracy.

- histogram – divide up range of $X: \Omega \to R'$ into k intervals: $c_0 < c_1 < \dots < c_k$

- Density histogram

★ $f_X^{dens}(x) = \sum_{z \in (c_i, c_{i+1})} \frac{f_X(z)}{(c_{i+1} - c_i)} = \frac{\#\{\omega : c_i < X(\omega) \le c_{i+1}\}}{\#(\Omega)(c_{i+1} - c_i)}$, when $x \in$

Then for $c_i > c_j$

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

- For such measurements we can introduce the idea of continuous approximation.

- let f be a probability density function fn so

(i) $f(x) \ge 0$

(ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

- then f approximation f_X^{dens} if $\forall\, a < b$

$$\int_a^b f(x)\,dx \approx \int_a^b f_X^{dens}(x)\,dx$$

make precise via limit as measurement are make to finer and finer accuracy.

- So continuous distribution (and infinite sample pass) arise as approximation

eg. Ω = students of U of T

- $X(\omega)$ = height in cm
- plausibly we cause the "approximation"

$X(\omega) \sim N(\mu, \sigma^2)$, $\forall \mu \in \mathbb{R}$, $\sigma^2 \in (0, \infty)$

- 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 U of T

- $Y(\omega)$ = weights in grams
- $X(\omega)$ = height in cms

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

$$\mu_y = \frac{1}{\#(\Omega)} \sum_{\omega \in \Omega} X(\omega) \quad , \quad \mu_x = \frac{1}{\#(\Omega)} \sum_{\omega \in \Omega} y(\omega)$$

$$\sigma_x^2 = \frac{1}{\#(\Omega)} \sum_{\omega \in \Omega} (x_i(\omega) - \mu_x)^2$$

$$\rho_{xy} = \frac{1}{\sigma_x \sigma_y} \frac{1}{\#(\Omega)} \sum_{\omega \in \Omega} (x_i(\omega) - \mu_x)(y(\omega) - \mu_x) = \frac{\sum_{\omega \in \Omega} (x_i(\omega) - \mu_x)(y(\omega) - \mu_y)}{\sqrt{\sum_{\omega \in \Omega} (x_i(\omega) - \mu_x)^2} \sqrt{\sum_{\omega \in \Omega} ((y(\omega) - \mu_y)^2)}}$$

- $y_i(\omega) \mid x(\omega) = x_2 \sim N\left(\mu_y + \sigma_y \rho_{xy} \frac{(x - \mu_y)}{\sigma_x}, \sigma_y^2 (1 - \rho_{xy}^2)\right)$
- So a relationship between X and y exists if $\rho_{xy} \neq 0$

- how strong is the relationship?
- the closer ρ_{xy}^2 is to 1 the stronger the stronger theoretically.
- note we can write

$$E(x_1 | x_2)(x_2) = \mu_1 + \frac{\sigma_1}{\sigma_2} \rho (x_2 - \mu_2)$$

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

$$= \beta_0 + \beta_1 x_2$$

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

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

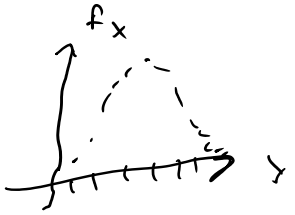
note - if we know f_x we know everything from a statistical point-of-view

note - when is a relationship between X & Y casual
- must be able to assign any value of X to any we ⊙

- Ω = population

$X : \Omega \to \mathcal{X}$ a measurement $\begin{pmatrix} x \\ y \end{pmatrix}$

for $x \in \mathcal{X}$ $f_X(x) = \dfrac{\#\{\omega : X(\omega) = x\}}{\#(\Omega)}$



- with a census we know everything

Sampling

- Typically we can't conduct 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 $x = (x_1, x_2 \ldots, x_n)$ we make inference about ?
- how do we select $\{\omega_1, \ldots, \omega_n\}$?
- clearly we want to avoid selection effects as this will "bias" our results/inferences.

eg. if $\omega \in \Omega$ = UofT students pass is always female then x would not be representative.

- Solution
  - we 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 the i-th number of Ω
- stir up the chips and pick one without looking and say the one labelled i.
- continue without replacement to obtain $x_1 = X(\omega_1), \ldots, x_n = X(\omega_n)$
- then select the population element

We model this via probability:

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

$$= f_X(x_1)$$

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

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

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

$$\approx f_X(x_2)$$

, given you removed the first person what's the probability of getting the second person.

, we can express it in terms of origins frequency function.

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

$$\approx f_X(x_1) f_{x_2}(x_2) \Rightarrow \perp \quad , \text{when } \#(\Omega) \text{ is large}$$

- Provided $n \ll \#(\Omega)$ (much smaller)

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

- So $x_1, \ldots, x_n$ are approximately 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 refering to the data as being objective

Statistical Models

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

eg. $A \subseteq X$

- $P(A) = \sum_{x \in A} f_X(x)$ = proportion of $\omega \in \Omega$ s.t $X(\omega) \in A$
  (of individuals in Ω)

- estimate by $\hat{P}(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(X(\omega_i))$

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

  SLLN

  $\rightarrow P(A)$ as $n \rightarrow \infty$

- $I_A(X(\omega)) \sim$ Bernoulli$(P(A))$, when ω is 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 $n\hat{P}(A) \sim$ Binomial$(n, P(A))$
- $\sqrt{n}(\hat{P}(A) - P(A)) \xrightarrow{D} N(0, \sqrt{\hat{P}(A)(1-\hat{P}(A))})$ CLT

... $P(A)$ with

- $\frac{\sqrt{n}\,(\hat{p}(A) - p(A))}{\sqrt{\hat{p}(A)(1-\hat{p}(A))}} \xrightarrow{d} N(0,1)$ CLT

- so $\hat{p}(A) \pm 3\sqrt{\frac{\hat{p}(A)(1-\hat{p}(A))}{n}}$ | contains $p(A)$ with virtual continuous

- but requires in "large"
- similarly for other characteristics of f_Y
  e.g μ_Y, σ^2_X, $\frac{\sigma_X}{\mu_Y}$, etc...
- want inference methods that don't require large n.
- Typically this involves making assumptions
  as sumption in value making (subjective) choices.

Principle of Empirical Criteria

- 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 for each $\theta \in \Theta$
- Θ is called the parameter of the model and Θ is called the parameter the parameter space.
- Typically Θ indexes; to each value of Θ there corresponds a unique density f_θ (no non identifiability)
- we have to check the assumption
  $f_X \in \{f_\theta : \theta \in \Theta\} = \mathcal{M}$

how? later

- So suppose we have checked $\mathcal{M}$ against x and have decided $\mathcal{M}$ is okay (note: in general we never say it is correct)
- then we want rules to apply to $(\mathcal{M}, x)$ to make inferences about true value of θ (and thus equivalent true f_Y)

Eg. Ω = students at U of T

- $X(\omega)$ = 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 standard residuals
- $\underset{\sim}{r} = \dfrac{\underset{\sim}{x} - \bar{x}\cdot}{\|\underset{\sim}{x} - \bar{x}\cdot\|} \overset{\text{fact}}{\sim} \text{Uniform}(S^{-1}(1) \wedge S^n)$

- propose test statistic that look for patterns

eg. skewness $T(\underset{\sim}{r}) = \frac{1}{n}\sum_{i=1}^{n} r_i^3 \overset{\text{SLLN}}{\to} 0$

- so should be close to zero

- Why make the assumption $f_x \in M$

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

(2) enables development of a theory of inference as opposed to asymptotes.

Types of Inferences

- Often we don't need to know θ but only $\psi = \Psi(\theta) \in \mathcal{I}$

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

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

- $\Psi(\mu, \sigma^2) = \mu$, inference about mean
- $\Psi(\mu, \sigma^2) = \sigma$, inference about standard deviation
- $\Psi(\mu, \sigma^2) = \frac{\sigma}{\mu}$
- $\Psi(\mu, \sigma^2) = \mu + \sigma z_{.75}$ (3rd quartile)
- $\Psi(\mu, \sigma^2) = \Phi\left(\frac{x_2 - \mu}{\sigma}\right) - \Phi\left(\frac{x_1 - \mu}{\sigma}\right)$

$=$ probability out of interval (x_1, x_2)

- two types of inferences (this is what we want the theory of inference to give)

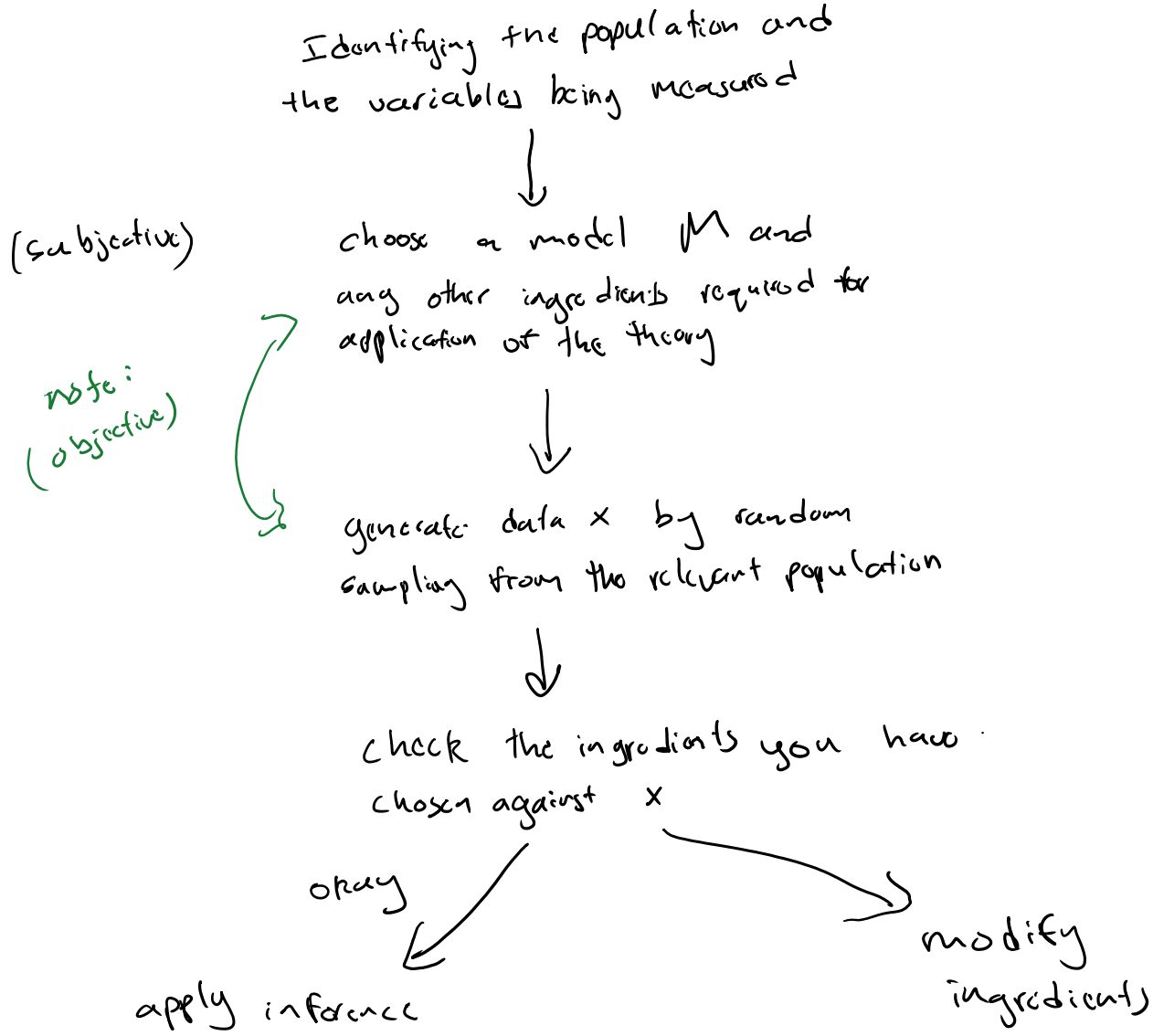
(1) Estimate $\hat{\psi}(x)$ together with an assessment of the accuracy of the estimate as given by a set $C(x) \subseteq \mathcal{I}$ with $\hat{\psi}(x) \in C(x)$ and size of $C(x)$ gives assessment.

(2) assess hypothesis $H_0 = \{\psi_0\} \subseteq \mathcal{I}$ by a quoting a measure of the evidence
- Basically means we have a hypothesis H_0 that the true value of ψ is $\psi_0 \in \mathcal{I}$ that H_0 is true together with a measure of the strength of this evidence.

- all theories of statistical inference attempt to censure 1 or 2.

Statistical Analysis

- We can conceive of a statistical analysis as the following steps.



Chapter 2

February 2, 2019 11:38 AM