where does the sampling dist. come from?

STAT 206
11 Feb 21
Lecture

$P \to (\mathcal{Q}, \mathbb{C}) \to$

$\to (\theta, D, B) \to M_{I,} = \{p(\theta|B), p(D|\theta B)\}$ ①

Bayesian model specification

where does the prior come from?

Mr.

$\to \ell(\theta|DB)$

Fisher didn't think he had uncertainty about:

"The model walks in the door with the client." (FALSE in general)

Mr. Fisher: $(\mathcal{Q}, \mathbb{C}) \to (\theta, D, B \to p(D|\theta B))$

---

usually context does not supply a unique choice for $p(\theta|B)$ or $p(D|\theta B)$

Bayesian statistical data science: ① build a probabilistic (BSDS) process model for data-generating process ② use Bayes' Thm. & its corollaries as optimal information-processing algorithm

① start with $\mathbb{P} = (\mathbb{Q}, \mathbb{C})$ & specify an algorithm that performs (eg., prediction) the task that answers $\mathbb{Q}$ well

Machine learning in data science

Disadvantages of BSDS: ① model uncertainty ② Bayesian inference & prediction may well not scale well

to really large (huge k, huge n, or both)

ML good at scaling their algorithms to Big Data size, but ML usually doesn't try to provide well-calibrated uncertainty assessments (this is a disadvantage of ML) (to be continued)

Bayesian vs. frequentist

---

statistical vs. ML machine learning data science

$$\underset{\sim}{I} = (I_1, I_2, \ldots, I_n, (\ldots))$$

binary → (arrows to I_1, I_2)

observed data vector

process of getting full-employment data

inference → θ = "true" pop. FE rate

$$\underset{\sim}{y} = (y_1, y_2, \ldots, y_n, (\ldots))$$

before data arrives

define H_1: start with

$$P(I_1 = y_1, \; I_2 = y_2, \ldots, I_n = y_n \mid \mathbb{B})$$

(and ↓ between terms; $\{0, 1\}$ ↓ above y_2; "data" circled, pointing to y_1)

prior predictive dist. for $\underset{\sim}{I}$

Quiz 2 case study

$n = 921$

inference: no feedback loop

prediction: feedback ✓

note: 2^n possible data vectors; $2^{921} \doteq 10^{277}$

if we know nothing about subjects

in the sample, our marginal predictive uncertainty about each person's response is the same! (the I_i are ID = marginally)

moreover, any logically-internally-consistent (coherent) predictive dist. must be invariant to the order in which data vector is given to us

invariant = remains the same under a transformation → group theory & other abstract algebra topics

exchangeability

123 { 123, 312, 132, 321, 213, 231 } all possible order permutations

exchangeability is a property of

your information about the world

↓ epistemology

not about the world itself

↓ ontology

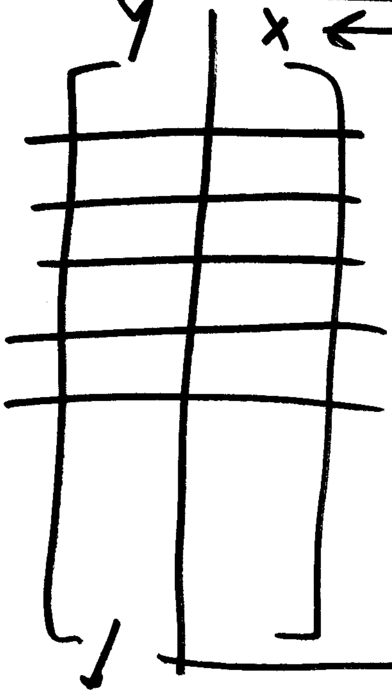
---

the judgment of exchangeability

comes from $\underline{\underline{B}} \leftarrow C$

↓ outcome

Y | X ← predictor / covariate / independent variable



n = 921

Quiz 2

case study

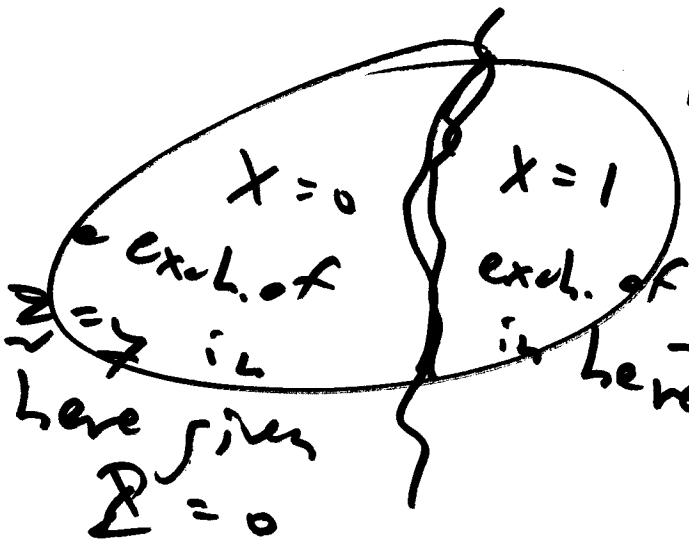
our uncertainty about $\underset{\sim}{X}$ is exch. if $\underset{\sim}{X}$ is not known

but if $\underset{\sim}{X}$ is known ⊛

$$y = \begin{cases} 1 & \text{if FE} \\ 0 & \text{not} \end{cases} \qquad x = \begin{cases} 1 & \text{if homeless} \\ 0 & \text{not} \end{cases}$$

our uncertainty about $\underset{\sim}{\Sigma}$ is no longer (unconditionally) exchangeable (partial) but becomes conditionally exch.

given $\underset{\sim}{X}$



De Finetti representation theorem for binary observables

(a) your uncertainty about $\underset{\sim}{\Sigma}=(\Sigma_1,\dots,\Sigma_n)$ is exch.

If (b) you're willing to extend π_n in (a) judgment of exch. from

$\underset{\sim}{\Sigma}=(\Sigma_1,\Sigma_2,\dots,\Sigma_n)$ (from sample) generalizing to

$\underset{\sim}{\Sigma}^*=(\Sigma_1,\Sigma_2,\dots,\Sigma_n,\dots)$ (to population) then

[1] (form $\bar{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} \Sigma_i$) then

$\lim_{n\to\infty} \bar{\Sigma}_n$ exists & $= \theta$ $(0 \le \theta \le 1)$

[2] $P(\Sigma_i = \gamma_i \mid \theta) = \begin{cases} \theta & \text{for } \gamma_i = 1 \\ 1-\theta & 0 \\ 0 & \text{else} \end{cases}$

(i.e.) $(\Sigma_i \mid \theta) \sim \text{Bernoulli}(\theta)$

and [3]

all coherent predictive dist. are expressible in following way:

$$P(\Sigma_1 = \gamma_1, \ldots, \Sigma_n = \gamma_n \mid \beta) = \int_0^1 \theta^s (1-\theta)^{n-s} \, p(\theta \mid \beta) \, d\theta$$

← likelihood from IID Bernoulli model

← prior pdf

where $s \triangleq \sum_{i=1}^{n} \gamma_i$

STAT 13

$$P(\underset{\sim}{Y} = y \mid \underset{=}{B}) = \int_0^1 P(\underset{\sim}{Y} = y, \theta \mid B)\, d\theta$$

(LTP)

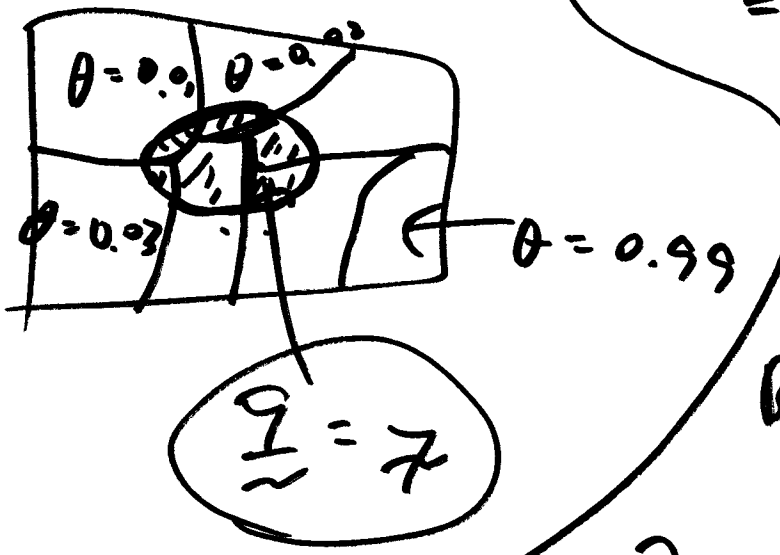
hard

but much easier if I knew θ ("truth")

$$= \int_0^1 \underbrace{P(\underset{\sim}{Y} = y \mid \theta, B)}_{f} \cdot \underbrace{p(\theta \mid B)}_{h}\, d\theta$$

$$= \int_0^1 \underbrace{\theta^s (1-\theta)^{n-s}}_{g} p(\theta \mid B)\, d\theta$$

$\Leftarrow$ (f-s)



Bayesian story: if λ is unknown to you, model your information/uncertainty about λ with a PDF/PMF $p(\lambda \mid \ldots)$, i.e. "treat λ as if it were a "random" variable

If your uncertainty about binary observables $(\Sigma_1, \ldots, \Sigma_n)$ not yet observed is exchangeable, the only coherent modeling must look like this:

$\theta \in (0,1)$

$$\left\{ \begin{array}{l} (\theta \mid \beta) \sim p(\theta \mid \beta) \quad \leftarrow \text{prior} \\ (\Sigma_i \mid \theta\, \beta) \overset{IID}{\sim} \text{Bernoulli}(\theta) \quad \leftarrow \text{freq.} \\ (i = 1, \ldots, n) \end{array} \right. \quad \text{Bayesian hierarchical model}$$

i.e. Bayesians in this model are frequentists with a prior

i.e. in this model, exch. $\rightarrow$ no sampling dist. uncertainty