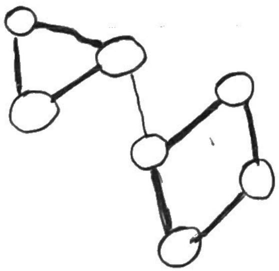
1/04/2020

NDAR dataset:
ADOS: 6.049 (NA omit 2,576)
ADI: 2,376 (NA omit 2,125)
MULLEN: 1,677 (NA omit 322)
SRS: 5,040 (NA omit 3,044)
VINELAND: 2,754 (NA omit 1,685)

726 (NA omit 0)   174 (NA omit 0)

- SRS → total_t_score
- MULLEN → exclude gm_t_score (vr, fm, rl, el)
- VINELAND → comm, livingskill, socialization, composite
- ADOS → A1, A2, B1, B4, A, B



q_{ij}

D1

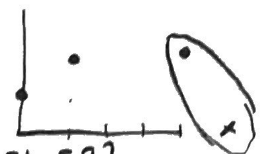
S1	1	2	3
S2	0	1	0
S3	4	2	1

D2

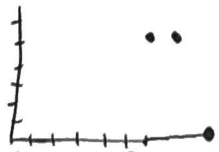
S1	6	5	6
S2	7	5	7
S3	8	0	1

D3

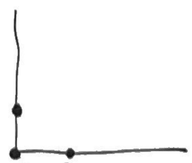
S1	1	1	0
S2	0	1	0
S3	0	0	1



{S1, S2}
{S3}



{S1, S2}
{S3}



{S1}
{S2}
{S3}

$$\begin{array}{c|ccc} & S1 & S2 & S3 \\ \hline S1 & 1 & 2/3 & 0 \\ S2 & 2/3 & 1 & 0 \\ S3 & 0 & 0 & 1 \end{array}$$

$S4_{D1}$: 0 5 ⇒ S3 (0%)
$S4_{D2}$: 6 5 ⇒ S1, S2 (2/3%)

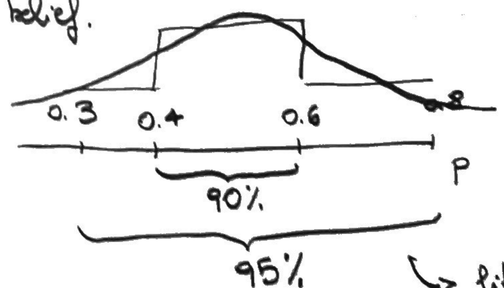
$\mathbb{P}(S_4 \in C_1 \mid \{S_1, S_2\}$ con $\mathbb{P} = 2/3)$

Bayesian statistics: "Introduction to B.S." book. MIT course Philippe Rigollet
- OBSERVE RANDOM DATA + PRIOR BELIEF
↓
• UPDATE TO POSTERIOR BELIEF after I see some data

EXAMPLE
p = proportion of women in pop.
Collect data $X_1, \dots, X_n \overset{iid}{\sim} Ber(p)$ * $p \in [0,1]$ 1 = woman, 0 = man
Frequentist approach: $\hat{p} = \frac{\sum X_i}{n} = \bar{X}_n$ → Test

We believe, in this case that $p \sim 0.5$. In Bayesian statistics we essentially consider 1/2 as a new observation. If we have 4 observation it will play a significant role, otherwise (with 1,000,000 obs) not so much.

We include p and then we update it into a posterior belief.



dist.
↓
param.
↓
data

↳ like in a Gaussian distribution

many ways we can have this. With a Gaussian we have a few n° of parameters (μ, σ).

Bayesian approach: way of modeling our belief about the parameter by doing <u>as if</u> it was random. For example we can say p is normally distributed. But we are on $[0,1] \Longrightarrow$

$$p \sim B(a,b) \quad \text{beta distribution}$$

$[0,1] \ni p$ — I want to find a distribution on $[0,1]$ (Gaussian is not OK).

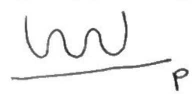
BETA DISTRIBUTION

$X \sim B(a,b) \qquad a,b>0$

$$f(x) = \begin{cases} x^{a-1}(1-x)^{b-1} & x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

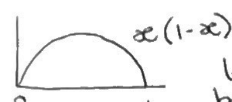
Unimodal distribution but twiking the parameters we can have a wide variety of shapes.

We want something where we might have some uncertainty but we are sure that it comes from one distribution and it is not like this



p is $\int f(x)\,dx$ (normalized)

$f(x) = c\,[x(1-x)]^{a-1}$



taking the power we distort this curve

$f(x) \propto [x(1-x)]^{a-1}$

we have endowed a parameter with a distribution

Now we draw p and then we generate data:

$$p \longrightarrow X_1,\dots,X_n \sim Ber(p)$$

The posterior distribution is still a beta distribution:

$$p \sim B\Big(\underbrace{a + \sum_{i=1}^{n} X_i}_{n^\circ \text{ of } 1s},\ \underbrace{a + n - \sum_{i=1}^{n} X_i}_{n^\circ \text{ of } 0s}\Big)$$

Since we have the same distribution as a prior and as a posterior $\rightarrow$ beta is conjugate prior (Also gaussian is a conjugate prior).

We want conjugate distributions.

To update the ~~post~~ prior we need the Bayesian theo.

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B)}$$

dist of data

dist of p

A = distribution of parameter given data

B = distribution of data given the parameter

Definitions: let $X_1,\dots,X_n$ be n random variables and θ parameter conditioned on.

$(X_1,\dots,X_n) \sim \underset{\text{pdf}}{P_n}(X_1,\dots,X_n|\theta)$

[example $(X_1,\dots,X_n) \sim N(\theta,1)$

$$p_n(X_1,\dots,X_n|\theta) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\sum_{i=1}^{n}\frac{(X_i-\theta)^2}{2}\right)$$

joint distribution]

We now assume

θ has prior distribution $\pi(\cdot)$ (either pdf, pmf)

[ex. $\pi(\theta) = C\theta^{a-1}(1-\theta)^{a-1}$

So now we have

$\mathbb{P}(B|A) = p_n(X_1,\dots,X_n|\theta)$

$\mathbb{P}(A) = \pi(\theta)$

$\mathbb{P}(B) = \int_\theta p_n(X_1,\dots,X_n)\,\pi(\theta)\,d\theta$ — margin distribution

Now we want the posterior:

$$\pi(\theta|X_1,\dots,X_n) = \frac{p_n(X_1,\dots,X_n|\theta)\,\pi(\theta)}{\int p_n(X_1,\dots,X_n|\theta)\,\pi(\theta)\,d\theta}$$

$f(x|y)$ conditional density, it relates to the $f(x,y)$ density and $f(x), f(y)$.

$$f_{X,Y}(x,y) = \underbrace{f_{X|Y=y}(x)}_{\text{conditional density}}\ \underbrace{f_Y(y)}_{\text{marginal pdf} = \int_{\mathbb{R}} f_{X,Y}(x,y)\,dx}$$

conditional density: $\dfrac{f_{X,Y}(x,y)}{f_Y(y)}$

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

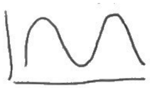
$$\underset{\text{POSTERIOR}}{\pi(\theta|X_1,\dots X_n)} = \frac{\boxed{p_n(X_1,\dots,X_n|\theta)}\ \overset{\text{PRIOR}}{\pi(\theta)}}{\int p_n(X_1,\dots X_n|\theta)\pi(\theta)\,d\theta}$$

// this does not depend on θ. ✱

Now we can compute: likelihood of θ

If the likelihood of θ is this 
and the prior of θ is this 

$\Downarrow$

the posterior is gonna be 

the posterior is the likelihood times something that puts some weight.
Bayesian estimation and maximum likelihood estimation are not that different.

$p \sim B(a,a)$

$\pi(p) \propto p^{a-1}(1-p)^{a-1}$

Likelihood (joint density of my data given p): $p_n(X_1,\dots,X_n|p) = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}$

MAP: maximum a posteriori

$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ \pi(\theta|X_1,\dots,X_n)$

$$p^{a-1+\sum X_i}(1-p)^{n-\sum X_i+a-1}$$

$$= p^{(a+\sum X_i)-1}(1-p)^{(a+n-\sum X_i)-1}$$

✱ methods that allow to sample from the posterior distribution without needing to compute the integral: MCMC
Monte Carlo Markov Chains.

we know it's a beta distribution, $\sim B(a+\sum X_i, a+n-\sum X_i)$
the normalization factor is known.

What if we face a new problem and we have no idea of the priors?
We need priors that carry no information (i.e., preferred parameters). So we need $\pi(\theta) \propto 1$ and we only look at the likelihood as a function we normalize in such a way that is actually a distribution.

These kind of priors are called UNINFORMATIVE.
Then the posterior becomes:

I uniform if Θ is bounded

II does not exist if Θ is not bounded

$$\pi(\theta|X_1,\dots,X_n) = \frac{\pi(X_1,\dots,X_n|\theta)}{\int_{-\infty}^{+\infty} p_n(X_1,\dots,X_n|\theta)\,d\theta}$$

if this converges we can do this.

Def (IMPROPER PRIOR): π is an improper prior on Θ if $\pi(\cdot)$ is a measurable nonnegative function defined on Θ not integrable.

So even if we have an IMPROPER PRIOR (case II) we still can compute posterior probabilities.

Let's say $p \sim \mathcal{U}([0,1])$

$$f(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{oth} \end{cases}$$

So then the posterior $p_n(p|X_1,\dots,X_n) = \dfrac{p^{\sum X_i}(1-p)^{n-\sum X_i}\cdot 1}{\int p^{\sum X_i}(1-p)^{n-\sum X_i}\,dp}$

$$B(1+\sum X_i, 1+n-\sum X_i)$$

$\mathbb{E}[X]$ with $X \sim B$?
It's not the avg.
If the n° of 1 is $0 \Rightarrow$ $1+\sum X_i = 1$ and we know $P(p=0|1)=0$
so we ~~avoid~~ include the "belief" that p is not always $=0|1$, although we don't have an informative prior.

The maximum of the posterior is the average of the X_i, i.e. the maximum likelihood estimator of a Bernoulli.

GAUSSIAN INFERENCE: (case improper prior)

$X_1,\dots,X_n \overset{iid}{\sim} \mathcal{N}(\theta,1)$ conditioned on θ

$$p_n(X_1,\dots,X_n|\theta) = \frac{1}{(\sqrt{2\pi})^n}\exp\left(-\frac{1}{2}\sum_{i=1}^n (X_i-\theta)^2\right)$$

$$\propto \exp(\dots)$$

and we see this as a density on θ.

$$= \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i^2 - 2X_i\theta + \theta^2)\right) \propto \exp\left(\theta\sum_{i=1}^{n}X_i - \frac{1}{2}\underbrace{\sum_{i=1}^{n}\theta^2}_{\frac{n}{2}\theta^2}\right)$$

$$-\frac{n}{2}\left[\underbrace{\theta^2 - \frac{2}{n}\sum X_i\theta}_{(\theta-\bar{X}_n)^2 - \bar{X}_n^2}\right]$$

we obtain $\Leftarrow$

$N(\bar{X}_n, 1/n)$

as a <u>posterior distribution</u>

$$\propto \exp\left(-\frac{n}{2}(\theta-\bar{X}_n)^2\right)$$

This is telling me that the more observation I get the more they will be concentrated around $\bar{X}_n$

<u>EXERCISE</u>: $\pi(\theta) = N(\mu, 1) \propto \exp\left(-\frac{1}{2}(\theta-\mu)^2\right)$ how do we updating the prior into a posterior? Prove this to be a conjugate prior, i.e. we have another gaussian an it is gonna look like we have another observation which is μ.

Prior list

- Beta prior
- Non informative prior
- Jeffreys prior: $\pi_J(\theta) \propto \sqrt{\det I(\theta)}$, where $I(\theta)$ is the Fisher information matrix. (GENERIC PRIOR as the uninf.).

So in this case we want to put weight on those θ that ~~maximize the~~ extract more information ~~we have~~ from the data.

let's say the model has the F. info. PROPORTIONAL TO THE AMOUNT OF INFORMATION THE MODEL HAS AT SOME POINT.

In previous examples we have (1 obs variance)

1. $\pi_J(p) \propto \frac{1}{\sqrt{p(1-p)}}$ $p \in (0,1)$ the prior is $\mathcal{B}(1/2, 1/2)$ "p closer to the boundary carry more info"
2. $\pi_J(p) \propto 1$, $\theta \in \mathbb{R}$ IMPROPER PRIOR (in the gaussian case). All θ carry the same amount of inf.

1.  $\rightarrow$ it pushes on the extreme values. Useful when we have $p = 1/2$ and we want to ~~...~~ force the model to make a decision (e.g. 0/1)

The $\pi_J(\theta)$ are invariant by the ~~domain~~ parametrization (e.g. p or p^2) of the function space can be reparametrized and still get Jeffrey's prior.

if we have p and we want to change it to p^2 we can.

2/04/2020

Def. (Fisher information)
Let $\{P_\theta\}_{\theta\in\Theta}$ denote a parametric family of distributions on a space X, with $\theta\in\Theta\subset\mathbb{R}^d$ and density p_θ. The Fisher information is the matrix
expectation taken wrt P_θ

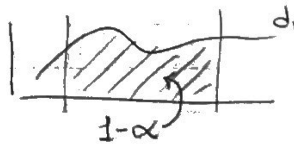
$$I_\theta = \mathbb{E}_\theta\Big[\underbrace{\nabla_\theta \log p_\theta(X)}_{\text{gradient of the likelihood at }\theta}\, \nabla_\theta \log p_\theta(X)^T\Big] = \mathbb{E}_\theta[\dot{l}_\theta \dot{l}_\theta^T]$$

variance of the likelihood estimator

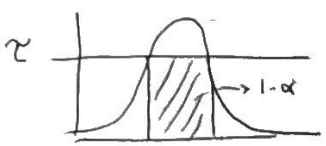
prior π, $X_1,\dots,X_n$ $\longrightarrow$ Bayes $\longrightarrow$ posterior distribution $\pi(\theta|X_1,\dots,X_n)$

Let's see how to build a BAYESIAN CONFIDENCE REGION (it can not be an int) (i.e. random subspace $R\subset\Theta$) conditioned on the data respect to posterior

$$\mathbb{P}(\theta\in R|X_1,\dots,X_n) = 1-\alpha \quad \text{at level } (1-\alpha)$$



(1) Level sets: let's say we have a gaussian posterior



threshold τ

if $\tau\longrightarrow+\infty \implies AUC\longrightarrow 0$
$\tau\longrightarrow 0 \implies AUC\longrightarrow 1$

If we have a bimodal posterior

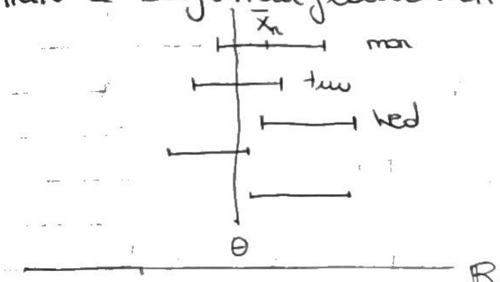
then we'll have union of confidence regions.

Let remind us of the frequentist CONF. INT.

$$P\Big(\theta\in\Big[\bar{X}_n \pm 1.96\frac{\sigma}{\sqrt{n}}\Big]\Big) = 0.95 \quad \text{gaussian.}$$

if I condition this P for $X_1,\dots,X_n \implies$ the $P = \begin{cases}0 & \theta\notin R\\ 1 & \theta\in R\end{cases}$
(deterministic conf. int, i.e. $\bar{X}_n$ is actually a number so either is true or not)

Within a Bayesian framework ~~if I keep building the CI~~



95% of the time is in (frequentist approach)

we don't have to repeat this, because $\mathbb{P}(\theta\in R|X_1,\dots,X_n)$
$\theta\in\pi(\theta)$ prior

- $n=2$ obs
$X_1 = \theta+\varepsilon_1$ $\quad$ ε_i iid
$X_2 = \theta+\varepsilon_2$ $\quad$ $\mathbb{P}(\varepsilon_i=\pm1)=1/2$

$\theta=\pm1 \implies$ we get 4 different pairs of observation $= \begin{cases}(\theta+1,\theta+1)\\(\theta-1,\theta-1)\end{cases} \neq \begin{cases}(\theta+1,\theta-1)\\(\theta-1,\theta+1)\end{cases}$

$$R = \begin{cases}\{X_1-1\} & \text{if } X_1=X_2\\ \dfrac{X_1+X_2}{2} & \text{if } X_1\neq X_2\end{cases}$$

CONF. REG (Freq) at level 75% — 3/4 of cases (we do not consider the case $(\theta-1,\theta-1)$)

$P(\theta\in R) = 0.75$

$\{X_1-1\}\cap X_1 = X_2$ $\quad = P(\theta = \theta+\varepsilon_1-1 \cap \varepsilon_1=\varepsilon_2) = P(\varepsilon_1=1 \cap \varepsilon_2=1) = 1/4$
($\varepsilon_1=1$, $\varepsilon_2=1$)

$\{\frac{X_1+X_2}{2}\}\cap X_1\neq X_2$ $\quad = P\Big(\frac{\varepsilon_1+\varepsilon_2}{2}=0 \cap \varepsilon_1\neq\varepsilon_2\Big) = P(\varepsilon_1=1\cap\varepsilon_2=-1) + P(\varepsilon_1=-1\cap\varepsilon_2=1) = 1/2$

$$\frac{1}{4}+\frac{1}{2} = 3/4$$

~~if $X_1\neq X_2$ we get the right θ~~

Bayesian CONF REG $\implies$ let assume π is a prior on θ with $\pi(j)>0\ \forall j\in\mathbb{Z}$ (weak assumption on my prior). My posterior is gonna be (let's build a likelihood)

$$\pi(X_1,X_2|\theta) = P(5,7|\theta) = P(\varepsilon_1=-1)\cdot P(\varepsilon_2=1) = \begin{cases}1/4 & \theta=6\\ 0 & \theta\neq 6\end{cases}$$

[let us assume $X_1=5$, $X_2=7$] $\implies \theta=6$ (avg(X_1,X_2))

My posterior

$$\frac{\pi(\theta)\overbrace{p(5,6|\theta)}^{0\ \forall\theta\neq 6}}{\underbrace{\sum_{t\in\mathbb{Z}}\pi(t)\underbrace{p(5,6|t)}_{=0\ \forall t\neq 6}}_{=\pi(6)\cdot 1/4}} \Rightarrow p(\theta|5,7)=\begin{cases}1 & \theta=6\\ 0 & \theta\neq 6\end{cases}$$

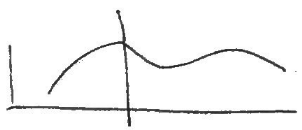
The confidence region here is equal to the value 1.

We can use frequentist methodology in Bayesian statistics. For instance we can use a bunch of priors and estimate the posterior, this estimate can then be tested.

Bayes estimators:

$$\hat{\theta}(\pi)=\int_{\Theta}\theta\,\pi(\theta|X_1,\dots,X_n)\,d\theta$$ (I) posterior mean

posterior median



posterior mode (MAP)

Example 1:

$X_1,\dots,X_n\sim Bern(p)$ $\Longrightarrow$ posterior

$p\sim Beta(a,a)$, $a>0$ $\quad \pi(p|X_1,\dots,X_n)\sim Beta(a+\sum X_i,\ a+n-\sum X_i)$

$p^{a-1}(1-p)^{b-1}$.

$$E[p]=\int_0^1 \underbrace{p\cdot\frac{p^{a-1}(1-p)^{b-1}}{C_{a,b}}\,dp}_{}$$ [posterior mean]

$$\frac{p^{a}(1-p)^{b-1}}{C_{a,b}}=\frac{C_{a+1,b}}{C_{a,b}}=\frac{a}{a+b}=\frac{a+\sum_{i=1}^n X_i}{2a+n}\xrightarrow[n\to+\infty]{}\bar{X}_n$$

the effect of the prior vanishes if $n\to\infty$

This is Jeffrey's prior $\left[\propto\sqrt{p(1-p)}=p^{1/2}(1-p)^{1/2}\right]$ when $a=1/2$.

If $n\sim 10\Longrightarrow$ depending on a the impact is visible.

Example 2:

When we have a non informative prior, i.e., $\pi(\theta)=1\ \forall\theta\in\mathbb{R}$ and $X_1,\dots,X_n\sim N(\theta,1)$ $\Longrightarrow \pi(\theta|X_1,\dots,X_n)\propto\exp\left(-\frac{1}{2}\sum_{i=1}^n(X_i-\theta)^2\right)$

$\sim N(\bar{X}_n, 1/n)$

↳ estimate posterior mean (just a point) we have more info from the posterior distribution

* The impact of the prior, if it's INFORMATIVE $\Longrightarrow$ it is washed away with $n\to+\infty$.

~~From the exercise of last time, with $n\to+\infty$~~

~~$N(\bar{\mu},\frac{1}{n+1})\xrightarrow[n\to+\infty]{}N(\bar{\mu},1)$~~

~~$\Rightarrow$ it is asymptotically normal~~

In both examples the Bayes estimator is asymptotically normal.