Lecture 1 2/1/16

Math ~~450~~ 380.03-02 KY 283

- Syllabus
- Review of 241

X is a non-deg. r.v. which means it has a support

$Supp[X] \subseteq \mathbb{R}$ s.t. $|Supp(X)| > 1$ i.e. more than 1 thing can be spit out

if $|Supp(X)| = |\mathbb{N}| \Rightarrow X$ is discrete i.e. its support is at most finite or ctbl. infinite

if $|Supp(X)| = |\mathbb{R}| \Rightarrow X$ is cont. i.e. its support is uncountably infinite and it spits out #'s which contain infinite information

e.g. time is considered a continuum which is infinitely divisible

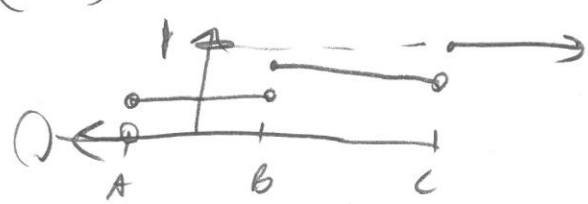
If discrete $\exists\ p(x)$ (PMF) which relates the support to a prob non-impossible, non-certain

(PDF) i.e. $p: Supp(X) \to (0,1)$

If cont. $\exists\ f(x)$ which represents the prob. density.

All r.v.'s have CDF's $F(x) := P(X \leq x) \Rightarrow P(X \in (a,b]) = F(b) - F(a)$

For discrete w/ support A, B, C, you get a step function



For continuous,

S.t. $f(x) := \frac{dF}{dx} = \lim_{\delta \to 0} \frac{F(x+\delta) - F(x)}{\delta} = \lim_{\delta \to 0} \frac{P(X \in [x, x+\delta])}{\delta}$

$\Rightarrow P(X \approx x) \approx \delta f(x) \propto f(x)$

so $\frac{P(X \approx x_1)}{P(X \approx x_2)} \approx \frac{f(x_1)}{f(x_2)}$ AKA a likelihood ratio

---

RV's are defined by their $\overbrace{\text{PMF/PDF/CDF}}^{\text{famous distr.}}$. Sometimes the ~~PMF/PDF/CDF~~ distr. is so common, we'll call them a brand-name r.v. e.g.

$X \sim \text{Bernoulli}(p) := p^x (1-p)^{1-x}$ $\quad Supp(X) = \{0, 1\}$

$X \sim \text{Binomial}(n, p) := \binom{n}{x} p^x (1-p)^{n-x}$ $\quad Supp(X) = \{0, 1, \ldots, n\}$

$X \sim \text{Exp}(\lambda) := \lambda e^{-\lambda x}$ $\quad Supp(X) = [0, \infty)$

$X \sim N(\mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ $\quad Supp(X) = \mathbb{R}$

$\vdots$

You must specify the "domain" of the PMF, PDF

But there are also these other pesky things:

$p^x (1-p)^{1-x}$ What is p?

Well if you remember the model, the Bernoulli gives $P(X=1)=p$, $P(X=0)=1-p$. So the only valid non-deg. values are $p \in (0,1)$ i.e. $\neq 0, \neq 1$.

p is a "parameter" of a "parametric model"

A ~~statistical~~ model attempts to assign probs for data using a stat. distr.

Data: Realis of r.v.'s x as opposed to X ← the r.v.
← the realization

Parameter Space: all poss. values of the parameter(s) of interest

From now on, all parameters are notated as θ
and the param space is $\theta \in \Theta$ ← capital theta in Greek

$\Rightarrow X \sim \text{Bern}(\theta) := \theta^x (1-\theta)^{1-x}$

$\Rightarrow X \sim \underbrace{N}(\theta_1, \theta_2^2) := \frac{1}{\sqrt{2\pi \theta_2^2}} e^{-\frac{1}{2\theta_2^2}(x-\theta_1)^2}$

$\vec{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$ is the vector of params

Note that $\dim(\vec{\theta}) = 2 < \infty$

Parametric Model: A set of distr's s.t. $\dim(\vec{\theta}) < \infty$ i.e. a family that can be "parametrized" with finite information.

$\mathcal{F} = \{ p(x;\theta) := \theta^x (1-\theta)^{1-x} : \theta \in \Theta \}$ ← s.t.

$|\mathcal{F}|$ is huge since $= |\Theta|$ but it's not that big.

$\mathcal{F}$ cannot model $\vec{X} = \langle 0, 1, 0, 1, 3.7 \rangle$ BLOW UP (will see)

ordered tuple notation

$$p(X;\theta)$$

"prob of X assuming θ"

or " " " " a param value of θ"

this means that prob changes depending on the param model picked within the family

sometimes this is $f(x;\vec{\theta})$ but we're not going to care...

$$\mathcal{F} = \{ p(x;\vec{\theta}) : \vec{\theta} \in \Theta \}$$ is the general model

usually $\dim(\vec{\theta}) = 1$ so we drop the vector notation

This class assumes (for the most part) the following problem:

$X_1, \dots, X_n$ are iid realizations from some model

independent, identically distributed

$$\Rightarrow p(X_1, \dots, X_n;\theta) = p(X_1;\theta) \cdot p(X_2;\theta) \cdot \ldots \cdot p(X_n;\theta) = \prod_{i=1}^{n} p(X_i;\theta)$$

AKA the "likelihood principle"

So in the real world you see $\langle 0, 0, 1, 0, 1, 0 \rangle$ AKA the "data"

$n=6$

Now, knowing something about how the data was generated, you pick an $\mathcal{F}$, a class of parametric models.

But you don't know θ!

Welcome to "inference" which broadly speaking has 3 goals:

1. Point Estimation. given $\vec{X}$, tell me the best guess of θ ~~[illegible]~~

2. Confidence set: give me a range of possible θ's based on ~~[illegible]~~ the "confidence level" I desire.

3. Hyp. Testing: I have an idea about what θ is. Does the data comport with this idea?

For example here, pick $\mathcal{F}$ to be the set of Bernoulli distr's.
Why? .... We know $\Theta \equiv (0,1)$ so θ could be anything in there.

$$p(\langle 0,0,1,0,1,0\rangle ; \theta = 0.5) = \prod_{i=1}^{6} p(x_i ; \theta = 0.5) = \prod_{i=1}^{6} 0.5^{x_i}(1-0.5)^{1-x_i} = 0.5^6 = 0.0156$$

$$'' \quad \quad '' \quad '' \quad \theta = 0.25 \; = \quad '' \quad '' \quad = 0.25^0 \, 0.75^1 \cdot 0.25^0 \, 0.75^1 \cdots$$

$$= 0.25^2 \, 0.75^4 = 0.0198$$

So $\theta = 0.25$ is "more likely" than $\theta = 0.5$.

How likely is θ?

$\mathcal{L}(\theta ; \vec{X})$ is the "likelihood" of θ assuming the data

$\stackrel{?}{=} p(\vec{X} ; \theta)$ Yes... it is the same function

but answers a different question...

the likelihood is a prob if F is discrete and a density if F cont.

What's the most likely value of θ?

$$\hat{\theta} := \underset{\theta \in \Theta}{\operatorname{argmax}}\ \mathcal{L}(\theta; x) = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \underbrace{\ln \mathcal{L}(\theta; x)}_{\text{log-likelihood}}$$

"maximum likelihood estimator" (MLE)

easier to make computations

Let's see

$\ln(a \cdot b) = \ln a + \ln b$

$$\ln \mathcal{L}(\theta; x) = \ln\left(\prod_{i=1}^{6} \theta^{x_i}(1-\theta)^{1-x_i}\right) = \sum_{i=1}^{6} \ln\left(\theta^{x_i}(1-\theta)^{1-x_i}\right)$$

$$= \sum_{i=1}^{6} x_i \ln(\theta) + (1-x_i)\ln(1-\theta)$$

$$= \ln\theta \sum x_i + \ln(1-\theta) \sum 1-x_i$$

Define $\bar{X} = \frac{1}{n}\sum x_i$ i.e. the sample average

$$= \ln(\theta)\, n\bar{X} + \ln(1-\theta)(n - n\bar{x}) \rightarrow n\ln(1-\theta)(1-\bar{x})$$

Now we need to find $\theta \in \Theta$ that max's this. So take deriv, set to 0.

$$\frac{d}{d\theta}\left[\ln \mathcal{L}(\theta; x)\right] = \frac{1}{\theta} n\bar{X} + (-1) n \frac{1}{1-\theta}(1-\bar{x}) = 0$$

$$\Rightarrow \frac{\bar{X}}{\theta} - \frac{1-\bar{X}}{1-\theta} = 0 \Rightarrow \frac{\bar{X}}{\theta} = \frac{1-\bar{X}}{1-\theta}$$

$$\Rightarrow \bar{x}(1-\theta) = \theta(1-\bar{x})$$

$$\Rightarrow \bar{x} - \cancel{\bar{x}\theta} = \theta - \cancel{\bar{x}\theta} \Rightarrow \theta = \boxed{\hat{\theta} = \bar{X}}$$

Should have done this in 241 $\bar{X} = \hat{p} = \frac{\cancel{\#}\#1s}{n}$