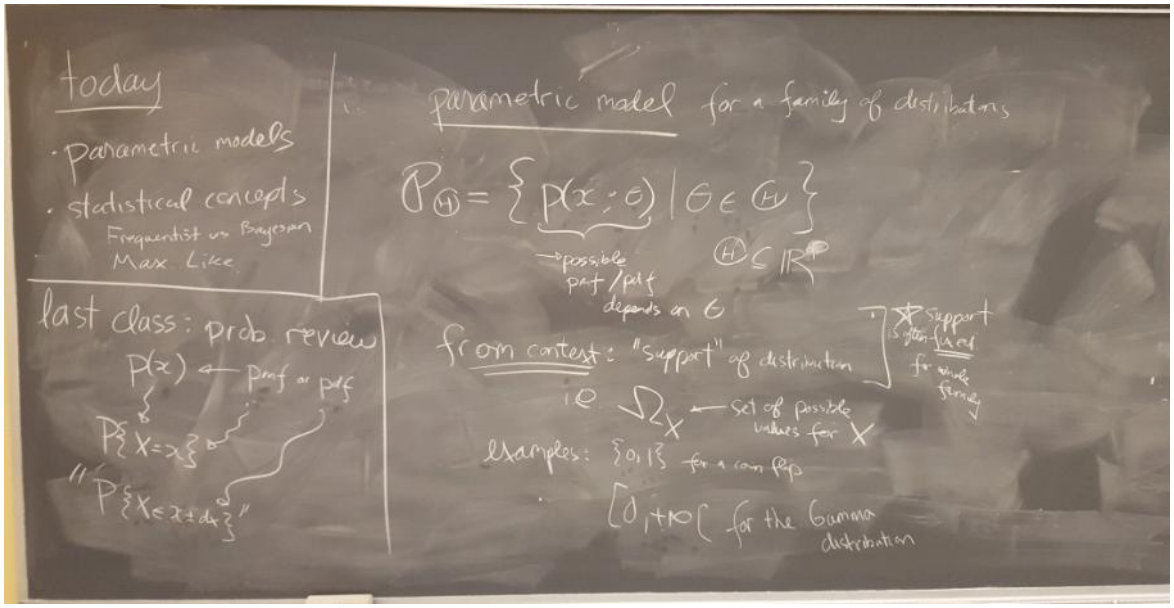
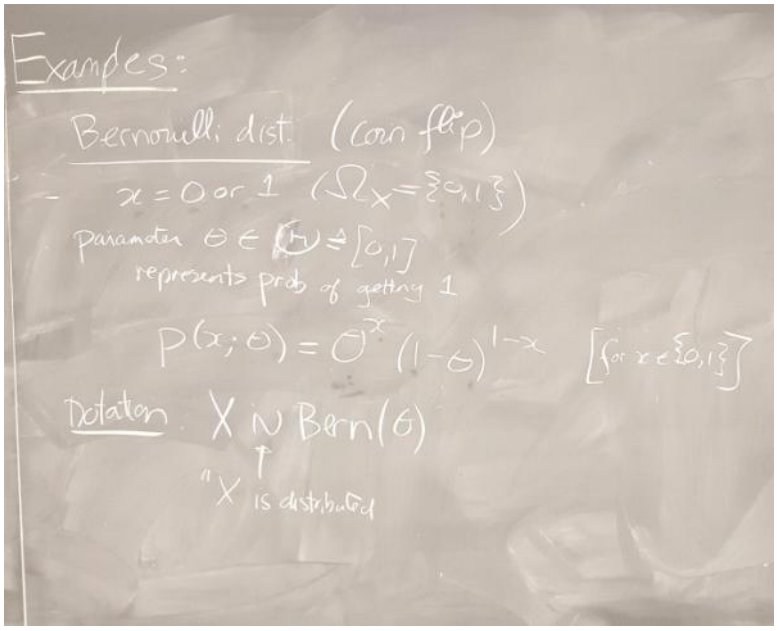
Lecture 3 - scribbles

Friday, September 9, 2016
13:35





notation: $X \sim \text{Bern}(\theta)$

"X is distributed as ..."

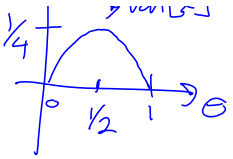
$p(x;\theta) = \text{Bern}(x;\theta)$ $\quad \text{Bern}(u;\theta)$ ...

variable for density, parameter

$\mathbb{E}[X] = \theta$

$\text{Var}[X] = \theta(1-\theta)$

$\mathbb{E}[X] = \theta$
$\mathrm{Var}[X] = \theta(1-\theta)$



- **binomial distribution**: → n independent coin flips
  → sum of n indep. Bern(θ) R.V. (often denoted p)

let $X_i \overset{\text{iid.}}{\sim} \text{Bern}(\theta)$ — independent & identically distributed

let $X = \sum_{i=1}^{n} X_i$ we have $X \sim \text{Bin}(n, \theta)$ (parameter; n: context)

$\Omega_X = \{0, 1, \dots, n\}$

$$p(x;\theta) = \underbrace{\binom{n}{x}}_{\substack{\#\text{ of ways to get } x \text{ successes} \\ \text{out of } n \text{ throws}}} \underbrace{\theta^x (1-\theta)^{n-x}}_{\theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)}} \quad \text{for } x \in \Omega_X$$

$$= \prod_{i=1}^{n} \text{Bern}(x_i;\theta)$$

(indep coin flips)

mean $X = \sum_i X_i$

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X_i] = n\theta \quad (\mathbb{E}[X_i] = \theta)$$

similarly $\mathrm{Var}[X] = n\theta(1-\theta)$

other distributions: Poisson(λ) (mean) $\Omega_X = \{0, 1, \dots\}$ [count data]

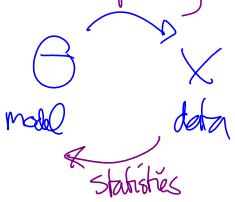
Gaussian (1D) $N(\mu, \sigma^2)$ (mean, variance) $\Omega_X = \mathbb{R}$

- gamma $\Gamma(\alpha, \beta)$ (shape, scale) $\Omega_X = \mathbb{R}_+$

others: Laplace, Cauchy, exponential, Beta & Dirichlet dist. .

Statistical concepts

cartoon



} well-defined

} ill-defined ("inverse problem")

epistemology of science

example: model of n indep coin flips

example: model of n indep coin flips

prob theory: prob of k heads in a row

statistics: I have observed n_0 tails, n_1 heads — what is θ?

Frequentist vs. Bayesian

philo: meaning of a probability??

a) (traditional) frequentist semantic

$p(X=x)$ represents the relative frequency of observing $X=x$

$\Rightarrow$ could repeat ∞-# of i.i.d. experiments

b) Bayesian (subjective) semantic

$p(X=x)$ encodes our "belief" that $X=x$

laws of probabilities characterizes a "rational" way to combine "beliefs" and "evidence"

[has motivation in terms gambling, utility/decision theory, etc...]

operationally:

- for a discrete R.V. suppose that $P(X=x)=\theta$

$$\Rightarrow P(X\neq x)=1-\theta$$

$$B=\mathbb{1}\{X=x\}\sim \text{Bern}(\theta) \text{ R.V.}$$

by L.L.N. $\frac{1}{n}\sum_{i=1}^{n} B_i \xrightarrow{a.s.} \mathbb{E}[B_i]=\theta$

↳ limiting frequency interpretation

[note: $\sum_{i=1}^{n} B_i \sim \text{Bin}(n,\theta)$ $\quad \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} B_i\right]=\frac{n\theta}{n}=\theta$

$$\text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} B_i\right]=\frac{1}{n^2}\text{Var}[\text{Bin}(n,\theta)]=\frac{1}{n^2}n\theta(1-\theta)$$

$$\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} B_i\right] = \frac{1}{n^2}\mathrm{Var}[\mathrm{Bin}(n,\theta)] = \frac{1}{n^2} n\theta(1-\theta)$$

(use $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$) $= \frac{1}{n}\theta(1-\theta)$]

$$\sqrt{n}\left(\frac{1}{n}\mathrm{Bin}(n,\theta) - \theta\right) \xrightarrow{d} N(0, \theta(1-\theta))$$

CLT central limit theorem

all works?

Bayesian approach

very simple philosophically: treat all uncertain quantities as R.V.

→ encode all your knowledge about the system ("beliefs") as a "prior" on probabilistic models

(later we'll see more with model selection)

simplest example: biased coin

we do not know θ

we believe $X \sim \mathrm{Bin}(n,\theta)$

→ θ is a R.V. for the Bayesian ⇒ need a $p(\theta)$ prior distribution

$\theta = [0,1]$

suppose observe $X = x$ (n results of flips)

then we can "update" our belief about θ using Bayes rule

$$p(\theta \mid X=x) = \frac{p(x\mid\theta)\,p(\theta)}{p(x)}$$

$p(\theta \mid X=x)$: posterior belief

$p(x\mid\theta)$: observation model / likelihood

$p(\theta)$: prior

$p(x)$: normalization

[note $p(x\mid\theta)$ → is pmf, $p(\theta)$ → is pdf

$p(x,\theta)$ is a mixed distribution]

example: suppose $p(\theta)$ is uniform on $[0,1]$ "no specific preference"

$$p(\theta \mid x) \propto \theta^x (1-\theta)^{n-x} \qquad [x \in \{0,1,\dots,n\}]$$

scaling: proportional

$$\int_0^1 \theta^x (1-\theta)^{n-x} d\theta = B(x+1, n-x+1)$$

$$B(a,b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \qquad \Gamma(a) \triangleq \int_0^\infty u^{a-1} e^{-u} du$$

Beta function

→ this is called a "Beta distribution"

$$\text{Beta}(\theta; \underbrace{\alpha, \beta}_{\text{parameters}}) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \mathbb{I}_{[0,1]}(\theta)$$

so $p(\theta|x) = \text{Beta}(\theta; x+1, n-x+1)$

look at $\mathbb{E}[\theta|x]$ (posterior mean)

as an estimator for θ

estimator function for θ

Notation: $\hat{\theta}(x)$ ← observation / "training data"

$$\hat{\theta}_{\text{Bayes}}(x) = \mathbb{E}[\theta|x] \qquad (\mathbb{E}[\theta|X=x])$$

mean for a $\text{Beta}(\alpha,\beta) = \frac{\alpha}{\alpha+\beta}$ so here $\mathbb{E}[\theta|x] = \frac{x+1}{n+2}$ ← this is result of experiment

constant → $n+2$

$$\hat{\theta}(X) = \mathbb{E}[\theta|X] = \frac{X+1}{n+2}$$

$$\hat{\theta} \sim \frac{1}{n+2}(\text{Bin}(n,\theta) + 1)$$

by LLN, if repeat m times independent experiment (with θ fixed)

we have $\frac{1}{m}\sum_{i=1}^{m} \hat{\theta}(X_i) \xrightarrow{a.s.} \mathbb{E}[\hat{\theta}] = \frac{n\theta+1}{n+2}$

$$= \left(\frac{n}{n+2}\right)\theta + \frac{2}{n+2}\left(\frac{1}{2}\right)$$

true parameter ← convex combination → prior mean

note that for finite n, $\mathbb{E}[\hat{\theta}(X)] \neq \theta$

→ "biased method" ← come from the prior belief

but as $n \to \infty$, $\mathbb{E}[\hat{\theta}_n(X)] \to \theta$

"asymptotically unbiased"

Beta(α,β) s.t. posterior mean of Beta($\alpha+x$, $\beta+n-x$) has mean of θ with respect to R.V. X

exercise?

Summary:

frequentist considers various "estimation" rules

[e.g. max. likelihood is most common]
but there are others
- method of moments
- M-estimation
- etc...

then they analyze "frequentist" properties of these estimators
- biased?
- variance?
- consistent?
- etc.

to get guarantees and compare them together

→ but no best rule [vs. Bayes rule for Bayesians]

sometimes MLE is "inadmissible" [e.g. James-Stein estimator]