$f(x;\theta)$ parametric estimation

a) Method of moments

$$\mu_i' = \int x^i f_X(x,\theta)\,dx = \mu_i(\theta)$$

- approximate (sample analog) the sample moment

$$\mu_1 = \mu_1(\theta) \qquad \bar{X} \xrightarrow{p} \mu$$

$$\mu_2' = \mu_2'(\theta) \qquad \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mu_2'$$

Let
$$\left.\begin{aligned} \mu_1(\hat\theta) &= \bar{X} \\ \mu_2'(\hat\theta) &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{aligned}\right\} \text{ solve for } \vec{\theta}$$

Example 7.2.1

$X_i \overset{iid}{\sim} N(\mu, \sigma^2)$

- $\hat\mu = \bar{X}$
- $\hat\mu_2' = \frac{1}{n}\sum_{i=1}^{n} X_i^2$

$$\Rightarrow \hat\sigma^2 = \hat\mu_2' - (\hat\mu)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = S^2$$

Remarks:

1. Robust and constructive
2. Estimators are functions of sums of rvs.
3. Resolts may be non-sensible

e.g. 7.2.2.

Let $X_i \overset{iid}{\sim}$ Binomial (k,p)

$$E[X_i] = kp \Rightarrow \hat{k}\hat{p} = \bar{X}$$

$$Var(X_i) = kp(1-p) \Rightarrow E[X_i^2] = kp(1-p) + k^2p^2$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \hat{k}\hat{p}(1-\hat{p}) + \hat{k}^2\hat{p}^2$$

$$\Rightarrow \hat{p} = \frac{\bar{X}}{\hat{k}}$$

$$\hat{k} = \frac{\bar{X}^2}{\bar{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

but $\hat{k}$ can be negative!

- MM is ignoring information here.

4. Many possible moments for most problems. (ie can use covariances.)
5. Getting "enough" moments is rarely the problem
6. Repopularized in economics (Hansen, EMA '82)
   - GMM

Maximum Likelihood

Defn 6.3.1 Let $f(x|\theta)$ be the joint pdf of the sample $X = (X_1, \ldots, X_n)$. Given that $X = x$ is observed, the function of θ given by $L(\theta|x) = f(x|\theta)$ is called the likelihood function.

e.g. Bin(n,p)

$$L(p|n,x) = \binom{n}{x} p^x (1-p)^{n-x}$$

<u>Defn 7.2.4</u>: For each sample point x, let $\hat{\theta}(x)$ be the parameter at which $L(\theta|x)$ achieves a maximum. A <u>maximum likelihood estimator</u> (MLE) of θ is $\hat{\theta}(x)$

e.g. $X_1, \dots, X_n \overset{iid}{\sim}$ Bernoulli(p)

$$L(p|x) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$= p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i} = p^{Y}(1-p)^{n-Y}$$

Taking first order conditions of $\log L = Y \log p + (n-Y)\log(1-p)$

(p): $\frac{Y}{p} = \frac{n-Y}{1-p} \Rightarrow Y - pY = np - pY$

$$\Rightarrow p = \frac{1}{n}Y = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}$$

<u>Remarks</u>:

1. Maximization is over the "parameter space": $\theta \in \Theta$ (e.g. $p \in [0,1]$)
   - can't get any "stupid" answers
2. Order preserving transforms
   ① Independence: $\prod_{i=1}^n f_X(x|\theta) \overset{\text{logs}}{\Rightarrow} \sum_{i=1}^n \log(f_X(x|\theta))$

② Many densities (esp. exponential family) have $e^{(\cdot)}$. Taking logs helps!

3] Main problem: ensuring maximum.

① Second-order conditions

② Local maximums

③ If Θ is bdd, watch out for boundary conditions

④ Multiple solutions. (identification)

4] "Pretend" normality $\Rightarrow$ MLE

- Quasi-MLEs (QMLE)

5] First order condition as function of θ from log-likelihood is the "score"

$$s(\theta|x) = \frac{\partial}{\partial\theta} \log L(\theta|x)$$

- Clearly, $s(\hat{\theta}|x) = \frac{\partial}{\partial\theta} \log L(\hat{\theta}|x) = 0$ if $\hat{\theta}$ is MLE.

6] Invariance properties.

- If $\hat{\theta}(x)$ is MLE of θ. Let T be a fcn of θ. Then $\widehat{T(\theta)} = T(\hat{\theta}(x))$

Bayes Estimators

$L(\theta|x)$

$\pi(\theta)$ density for θ reflecting uncertainty over θ

- the "prior"

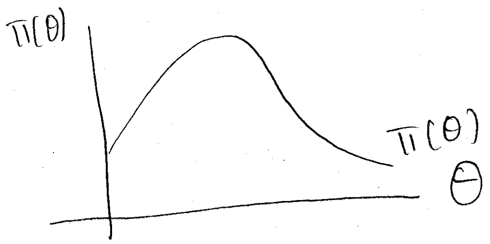
- $\pi(\theta)$ is what I thought θ might look like before I saw the data. What do I think after I see x?

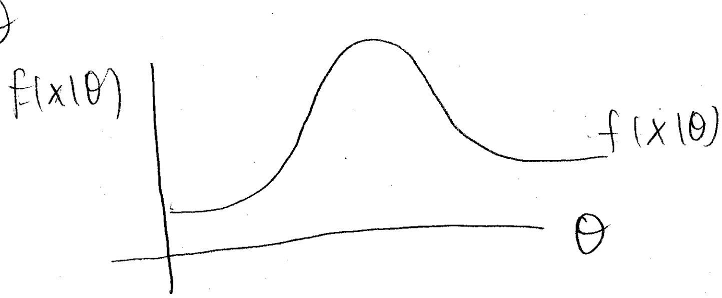
$$L(\mathbb{X},\theta) = \pi(\theta|x)\cdot \overset{\swarrow \text{ marginal of } \mathbb{X}.}{m(x)}$$
$$= f(\mathbb{X}|\theta)\cdot \pi(\theta)$$

$\Rightarrow f(\mathbb{X}|\theta)\cdot\pi(\theta) = \pi(\theta|\mathbb{X})\, m(\mathbb{X})$

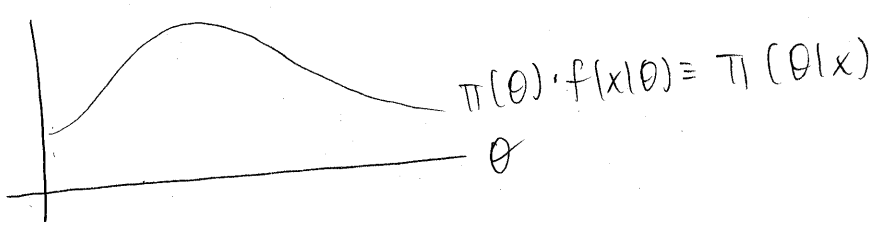
$\Rightarrow \underbrace{\pi(\theta|\mathbb{X})}_{\text{posterior density for } \theta} = \dfrac{f(\mathbb{X}|\theta)\,\pi(\theta)}{\underbrace{m(\mathbb{X})}_{\text{ensures we have a density (scales up/down)}}}$ "Bayes' rule"

$m(\mathbb{X}) = \int f(x|\theta)\pi(\theta)d\theta$





$\Rightarrow$



e.g. 7.2.14 $\mathbb{X}_i \overset{iid}{\sim}$ Bernoulli $\mathbb{Y} = \sum_{i=1}^{n} \mathbb{X}_i \sim Bin(n,p)$

Suppose $\pi(p) \sim Beta(\alpha, \beta)$

$m_{\mathbb{Y}}(y) = \int f(y|p)\pi(p)dp$

$f(y|p) = \binom{n}{y} p^y (1-p)^{n-y}$ $\pi(p) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$

$$\Rightarrow m_{\underline{Y}}(y) = \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\,\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}$$

$$\Rightarrow \pi(p|y) = \frac{f(y|p)\,\pi(p)}{m_{\underline{Y}}(y)} \sim \text{Beta}(y+\alpha, n-y+\beta)$$

$$E[\pi(p|y)] = \frac{y+\alpha}{y+\alpha+(n-y+\beta)} = \frac{y+\alpha}{\alpha+\beta+n} = \left(\frac{n}{\alpha+\beta+n}\right)\left(\frac{y}{n}\right) + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right)$$

- weighted average of sample mean and prior mean.
- "a Bayesian is a person who expects to see a horse, sees a donkey, and concludes it's a mule."

Remarks:

1. Precise mathematical formulation to "what do I learn when I observe $\underline{X}=x$?"
2. Results depend on prior.
3. Where does $\pi(\theta)$ come from?
   - mathematical tractibility (conjugate priors)
     - conjugate of normal is normal
     - conjugate of binomial is beta
4. Improper priors: $\int \pi(\theta)\,d(\theta) \neq 1$
   - e.g. $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$

     $\propto$ is proportional to

- can force it to integrate to one.

$\boxed{5}$ often, certain priors lead to classical results (MLEs)

<u>Next time</u>: Evaluating these estimators.