Today: Notes 3, get as far as we can

Monday: Lab, wrap up Notes 3 → Lab due Wed night
HW due Wed. night,
graded on completion
I'll post solutions on Tues

Wednesday: Quiz

Tuesday: Stat speaker @ 4:15/4:30
→ EC opportunity!

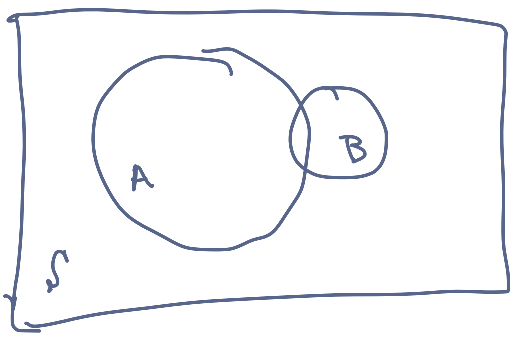
03: BAYESIAN ESTIMATION

Larsen & Marx 5.8

Prof Amanda Luby

1 Bayes Theorem

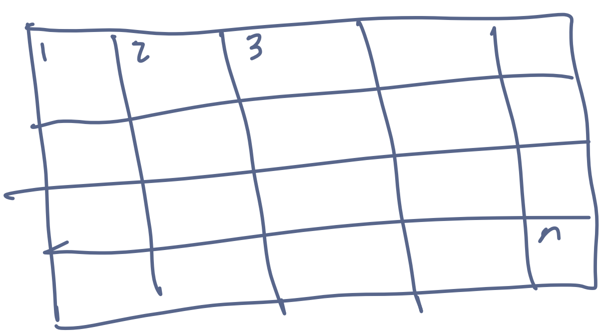
Idea: if you know $P(A|B)$, how can you find $P(B|A)$?
"inverse probability"



$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

Bayesian statistics is a set of techniques that are based on inverse probabilities calculated using Bayes' theorem. Relative to "classical techniques" (MoM and MLE), Bayesian estimation provides a way to incorporate "prior knowledge" into the estimation of parameters.

Example: 1968 submarine went missing USS Scorpion



A_1: sub sunk in sec 1
⋮
A_n: sub sunk in sec n

Solicited $P(A_1) \dots P(A_n)$ from experts

1. Idea: pick largest A_i call A_k and search that one first

B_k: sub would be found in k if k was searched
- function of water depth

B_k^c: k was searched & sub not found

2. $$P(A_k|B_k^c) = \frac{P(B_k^c|A_k)P(A_k)}{P(B_k^c|A_k)P(A_k) + P(B_k^c|A_k^c)P(A_k^c)}$$

3. P(sub sank in k | not found in k) becomes "updated" $P(A_k) \to P^*(A_k)$

4. renormalize $P(A_j)$ for $j \neq k \to P^*(A_j)$
search largest $P^*(A_j)$ and repeat
$P^{**}(A_j)$ etc...

2 key pieces:

1. incorporate "prior knowledge"

2. Mechanism to update $P(A_k)$ with new information

Classical Statistics	Bayesian Statistics
Probability refers to limiting relative frequencies. Probabilities are objective properties of the real world.	*Probability* describes a degree of belief, not a limiting frequency. As such, we can make probability statements about lots of things, not just data which are subject to random variation. For example, I might say that "the probability that Albert Einstein drank a cup of tea on August 1, 1948 is .35". This does not refer to any limiting frequency. It reflects my strength of belief that the proposition is true.
Parameters are fixed, unknown constants, and the data we observe is random. Because they are constant, no useful probability statements can be made about parameters.	*Parameters* are random, and the data that we observe are fixed. We can therefore make probability statements about parameters.
Statistical procedures should be designed to have well-defined long-run frequency properties. For example, a 95% confidence interval should capture the true value of the parameter at least 95% of the time.	We make inferences about a parameter θ by producing a probability distribution for θ. Inferences, such as point estimates and interval estimates, may then be extracted from this distribution.

Bayesian inference is a controversial approach because it inherently embraces a subjective notion of probability. The field of statistics generally puts more emphasis on frequentist methods although Bayesian methods definitely have a presence.

2 Bayesian Inference

1. **Prior distribution:** $f_\theta(\theta)$ $\quad$ $P_\theta(\theta)$ if discrete

   degree of belief about θ before we see any data

   sub example: $P(A_k)$'s

2. **Statistical model for data:**

   $f_X(X|\theta)$: belief about the data given a parameter θ

   NOTE: $f_X(X;\theta)$ is different than $f_X(X|\theta)$

   sub example: $P(B_k)$'s

3. **Posterior distribution:**

   $f_{\theta|X}(\theta|X)$: updated belief about θ after seeing our data

ex: $P(A_k|B_k^c) \to p^*$

If we see $w_1, \ldots, w_n$ replace $P_W(w|\theta)$ with $\prod_{i=1}^{n} P_W(w_i|\theta) = L(\theta, w)$

Posterior distribution

Let W be a statistic dependent on parameter θ. Call its pdf $f_W(w|\theta)$. Assume that θ is the value of a random variable Θ, whose prior distribution is denoted p_Θ if discrete and f_Θ if continuous. The *posterior distribution* of Θ given $W = w$ is:

$$f_{\Theta|W} = \begin{cases} \dfrac{P_W(w|\theta) f_\Theta(\theta)}{\int_{-\infty}^{\infty} P_W(w|\theta) f_\Theta(\theta)\, d\theta} & w \text{ discrete} \\[2ex] \dfrac{f_W(w|\theta) f_\Theta(\theta)}{\int_{-\infty}^{\infty} f_W(w|\theta) f_\Theta(\theta)\, d\theta} & w \text{ continuous} \end{cases}$$

If θ is discrete, replace integrals w/ sums and f_Θ with p_Θ

4. **Posterior mean**: Estimator: $\hat{\theta} = E(\theta | w)$

$$= \int_{-\infty}^{\infty} \theta \cdot f_{\theta|w}(\theta|w)\, d\theta$$

Example: Let $X_1, \ldots, X_n \sim$ Bernoulli($\cancel{p}\ \theta$) and suppose that $\cancel{p}\ \theta$ has the prior distribution $\cancel{p}\ \theta \sim \text{Beta}(\alpha, \beta)$.

$P(X_i = x) = \theta^x (1-\theta)^{1-x} \quad x = \{0,1\}$

$f_\theta = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad 0 \le \theta \le 1$

let $X = \sum X_i \quad X \sim \text{Bin}(n, \theta)$

$P(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$

Goal: find posterior distribution of $\theta | X$: $\dfrac{P_X(x|\theta) f_\theta(\theta)}{\int P_X(x|\theta) f_\theta(\theta)\, d\theta}$

numerator: $P_X(x|\theta) f_\theta(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \cdot \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$

$$= \underbrace{\binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{\text{factor constant out of top \& bottom}} \underbrace{\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}_{\text{"kernel" of a Beta pdf}}$$

make denominator pdf of Beta$(x+\alpha, n-x+\beta)$

$$f_{\theta|x} = \frac{\cancel{\binom{n}{x}\Gamma(\alpha+\beta)}\big/\cancel{\Gamma(\alpha)\Gamma(\beta)}}{\cancel{\binom{n}{x}\Gamma(\alpha+\beta)}\big/\cancel{\Gamma(\alpha)\Gamma(\beta)}} \cdot \frac{\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{\int \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}\, d\theta} = \frac{\frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{\int \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}\, d\theta}$$

$\Rightarrow$ Posterior is a Beta$(x+\alpha, n-x+\beta)$

$$f_{\theta|x} = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

What we did after recognizing the kernel of a Beta RV is mess w/ normalizing constant.
If we recognize kernel, can always factor constant out of numerator + denominator, then multiply by a more useful constant in both numerator and denominator.

$\Rightarrow$ we don't have to go through the trouble if we recognize the kernel

Shortcut : $f_{\theta|X} \propto f_{X|\theta}(x|\theta)\, f_\theta(\theta)$

$$\propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

"proportional to"

"up to a normalization constant"

Our <u>Bayes Estimator</u> is theoretical mean of posterior

$E(\theta|X)$

for Beta: $E(\theta|X) : \dfrac{x+\alpha}{n-x+\beta+x+\alpha} = \dfrac{x+\alpha}{n+\beta+\alpha}$

Today:

- Wrap up Notes 3
- Overview of Quiz Expectations
- Lab 02

- Quiz 1 on Wed
- HW 3 & lab 2 due Wed night
- Post solutions on Tues
- Colloquium speaker @ 4:15
- Moving Wed OH to Tuesday 2:30-4

Recap

posterior distribution:

$$f_{\theta|X} \propto f_{X|\theta}(x|\theta)\, f_\theta(\theta)$$

If $X_1, \dots, X_n$ is our data, replace $f(x|\theta)$ with $f(x_1, \dots, x_n) = \prod f(x_i|\theta) = L(x_i|\theta)$

Then

\propto

$$f(\theta|x^n) = \frac{f(x^n|\theta)\, f(\theta)}{\int f(x^n|\theta)\, f(\theta)\, d\theta} = \frac{L_n(\theta) f(\theta)}{\underbrace{\int L_n(\theta) f(\theta)\, d\theta}_{\text{constant that does not depend on } \theta}} \propto L_n(\theta) f(\theta)$$

Conjugate prior

When the prior and posterior are in the same family of distributions (same name) we say the prior is <u>conjugate</u> for that likelihood

Ex: Beta is the conjugate prior for binomial likelihood

Example: Let $X_1, ..., X_n \sim \underbrace{N(\theta, \sigma^2)}_{\text{parameters}}$ and suppose we take $\theta \sim \underbrace{N(a, b^2)}_{\text{hyper parameters}}$. For simplicity, let's assume σ^2 is known.

Goal: find posterior distribution $\theta | X^n$

$$f_{\theta|X}(\theta|x) \propto f_{X^n}(x|\theta) f_\theta(\theta) = \left[\prod \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i-\theta)^2}{2\sigma^2}\right)\right] \cdot \frac{1}{\sqrt{2\pi b^2}} \exp\left[\frac{(\theta-a)^2}{2b^2}\right]$$

$$\propto \exp\left(-\frac{1}{2\sigma^2}\sum(x_i-\theta)^2\right) \exp\left(-\frac{1}{2b^2}(\theta-a)^2\right)$$

$$= \exp\left(-\frac{1}{2\sigma^2}\sum(x_i-\theta)^2 - \frac{1}{2b^2}(\theta-a)^2\right)$$

$$= \exp\left(-\frac{\sum x_i^2}{2\sigma^2} + \frac{2\theta\sum x_i}{2\sigma^2} - \frac{n\theta^2}{2\sigma^2} - \frac{\theta^2}{2b^2} + \frac{2a\theta}{2b^2} - \frac{a^2}{2b^2}\right)$$

$$= \exp\left(\theta^2\left(\frac{-n}{2\sigma^2} - \frac{1}{2b^2}\right) + \theta\left(\frac{2\sum x_i}{2\sigma^2} + \frac{2a}{2b^2}\right) + \left(\frac{\sum x_i^2}{2\sigma^2} - \frac{a^2}{2b^2}\right)\right)$$

From here, want to show the above expression can be written as

$$\exp\left(-\frac{1}{2\sigma_*^2}(\theta-\theta_*)^2\right) = \exp\left(-\frac{1}{2\sigma_*^2}\left(\theta^2 - 2\theta\theta_* + \theta_*^2\right)\right)$$

Equate like terms:

$$\frac{-\theta^2}{2\sigma_*^2} = -\theta^2\left(\frac{n}{2\sigma^2} + \frac{1}{2b^2}\right)$$

$$\frac{1}{\sigma_*^2} = \left(\frac{nb^2 + \sigma^2}{\sigma^2 b^2}\right)$$

$$\boxed{\sigma_*^2 = \frac{\sigma^2 b^2}{nb^2 + \sigma^2}}$$

$$\frac{2\theta\theta_*}{2\sigma_*^2} = \theta\left(\frac{\sum x_i}{\sigma^2} + \frac{a}{b^2}\right)$$

$$\frac{\theta_*}{\sigma_*^2} = \frac{\sum x_i}{\sigma^2} + \frac{a}{b^2}$$

$$\theta_* = \sigma_*^2\left(\frac{\sum x_i}{\sigma^2} + \frac{a}{b^2}\right)$$

$$\theta_* = \frac{\sigma^2 b^2 \sum x_i}{\sigma^2} + \frac{\sigma^2 b^2 a}{(nb^2+\sigma^2)b^2} = \frac{b^2\sum x_i}{nb^2+\sigma^2} + \frac{\sigma^2 a}{nb^2+\sigma^2} = \boxed{\frac{b^2\sum x_i + \sigma^2 a}{nb^2+\sigma^2}}$$

$$\Rightarrow \theta|X \sim N(\theta_*, \sigma_*^2) = \boxed{N\left(\frac{b^2\sum x_i + \sigma^2 a}{nb^2+\sigma^2}, \frac{\sigma^2 b^2}{nb^2+\sigma^2}\right)}$$

bayes estimator: $E(\theta|X) = \frac{b^2}{b^2+\sigma^2/n}\bar{X} + \frac{\sigma^2/n}{b^2+\sigma^2/n} \cdot a$

weighted average of sample mean (also MLE) $\bar{X}$ and prior mean a. As $n \to \infty$, $E(\theta|X) \to \bar{X}$

Quiz 1

Quiz:

~30 minutes

mix of concepts and mechanics

OH today 11:30–12:30

tues 2:30–4

- Terminology
  - parameter vs estimator
  - estimator vs estimate
  - pdf vs likelihood
  - sample vs population
  - prior vs posterior
- Basic Integration
  - polynomial: $x^2 + x + c$
  - e^x
  - $\ln(x)$
  - simple chain rules $\int \ln(2x)$
- Given a pdf, find Expected value + variance
  - recognize named distibutions
  - use properties of expected value + variance
  - simple integration
- Given $E(\hat{\theta})$ and $V(\hat{\theta})$
  - comment on bias + efficiency
- Estimation
  - MLE - set up + find
  - MoM
  - Bayes Estimator: multiply $L_n(\theta) \cdot f_0(\theta)$
    use kernel to recognize named posterior