Bayesian Inference, continued

STAT 244
1/26/16
Lecture 7

In the example of polling from last time, there is in fact some prior knowledge - for example, polls in other states show Clinton and Sanders close to even. We have the notion that Illinois Democrats are not that different than Democrats in other states.

How to make use of this knowledge? With a <u>richer</u> <u>class</u> <u>of</u> <u>priors</u>:

We took f_Y as Beta, so

$$f_Y(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

(remember: the Uniform dist is a special case of the Beta dist when $\alpha = \beta = 1$).

It turns out that as α and β get bigger, and are not too different, the Beta distribution begins to resemble a Normal dist.

For $N(\mu, \sigma)$,

$$P(|Y-\mu| < \sigma) \approx 2/3 \approx .667$$

For Beta,

α	β	$P(\lvert Y-\mu \rvert < \sigma)$
1	1	.577
2	2	.626
3	3	.644
5	5	.659
4	6	.661
1	19	.812 ← SKEWED!
10	10	.671

... etc.

So suppose we expect that the vote is approximately split, with $Pr(.4 < Y < .6) \approx 2/3$.

So

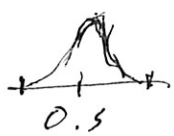
$$E(Y) \cong 0.5$$

$$Var(Y) \cong 0.1$$

For Beta distributions,

$$E(Y) = \frac{\alpha}{\alpha+\beta}, \quad Var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

set these equal to 0.5 and $(0.1)^2$, solve to get $\alpha = \beta = 12$



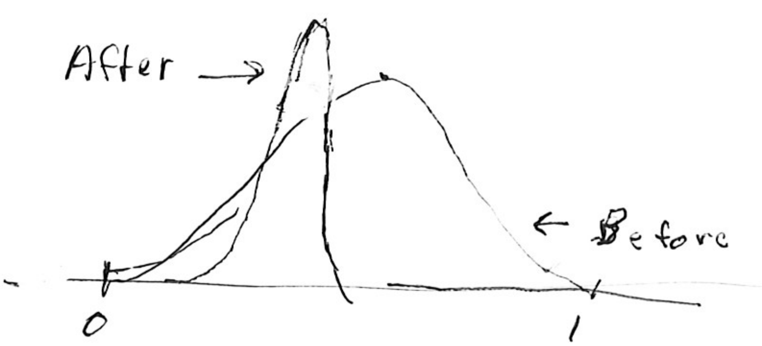
$$f_Y(y) = \frac{\Gamma(24)}{\Gamma(12)\Gamma(12)} y^{11}(1-y)^{11}$$

has $E(Y) = 0.5$, $\sqrt{Var(Y)} = 0.1$

$$f(y|x) \propto f_Y(y)\, p(x|y)$$

$$= (\text{constant})\, y^{11}(1-y)^{11}\binom{100}{40} y^{40}(1-y)^{60}$$

$$\propto y^{51}(1-y)^{71}$$



[Beta with $\alpha = 52$, $\beta = 72$]

Before: $E(Y) = 0.5$

After: $E(Y|X=40) = \frac{52}{124} = \underline{\underline{0.42}}$

For Beta distributions,

$$E(Y) = \frac{\alpha}{\alpha+\beta} = \mu_Y$$

$$Var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \sigma_Y^2$$

$$\sigma_Y^2 = \overbrace{\left(\frac{\alpha}{(\alpha+\beta)}\right)}^{=\mu_Y}\left(\frac{\beta}{(\alpha+\beta)}\right)\frac{1}{(\alpha+\beta+1)}$$

$$\frac{\beta}{\alpha+\beta} = \frac{\beta+\alpha-\alpha}{\alpha+\beta} = \frac{\beta+\alpha}{\alpha+\beta} - \frac{\alpha}{\alpha+\beta}$$

$$= 1 - \mu_Y$$

so

$$Var(Y) = \frac{\mu_Y(1-\mu_Y)}{\alpha+\beta-1}$$

$$\mu_Y = 0.5 \quad \sigma_Y^2 = (0.1)^2$$

So:

$$(0.1)^2 = \frac{(0.5)(1-0.5)}{\alpha+\beta+1}$$

$$\alpha+\beta+1 = \frac{(0.5)^2}{(0.1)^2} = 25; \quad \alpha+\beta = 24$$

but

$$\mu_Y = \frac{\alpha}{\alpha+\beta} = 0.5 \Rightarrow \frac{\alpha}{24} = 0.5 \Rightarrow \alpha = 12, \beta = 12$$

Note: We can interpret the posterior expectation as a weighted average:

$$\frac{\alpha + x}{\alpha + \beta + n} = \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right) \cdot \frac{\alpha}{\alpha + \beta} + \left(\frac{n}{\alpha + \beta + n}\right) \cdot \frac{x}{n}$$

weights add to 1

prior expectation

sample fraction

Case 1: $\alpha + \beta$ large relative to n ("strong prior information")

Then $\frac{\alpha + \beta}{\alpha + \beta + n} \approx 1$, $\frac{n}{\alpha + \beta + n} \approx 0$

Case 2: n large relative to $\alpha + \beta$ ("weak prior information")

Then $\frac{\alpha + \beta}{\alpha + \beta + n} \approx 0$, $\frac{n}{\alpha + \beta + n} \approx 1$

Case 1: $\frac{\alpha}{\alpha + \beta}$ case 2: $\frac{x}{n}$

otherwise, a compromise!

Bayes for Normal

θ is the true value.

We take the prior $f(\theta)$

$$f(\theta) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{\theta-\mu}{\sigma}\right)^2}, \quad -\infty < \theta < \infty$$

$$E(\theta) = \mu \qquad Var(\theta) = \sigma^2$$

X is the observed value

with error $\sim N(0, \tau^2)$ so

$X = \theta$ + error is $N(\theta, \tau^2)$

$$\underbrace{f(x|\theta) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\theta}{\tau}\right)^2}}_{\text{likelihood}} \quad -\infty < x < \infty$$

$f(\theta|x)$, the posterior will be $N(A, B^2)$

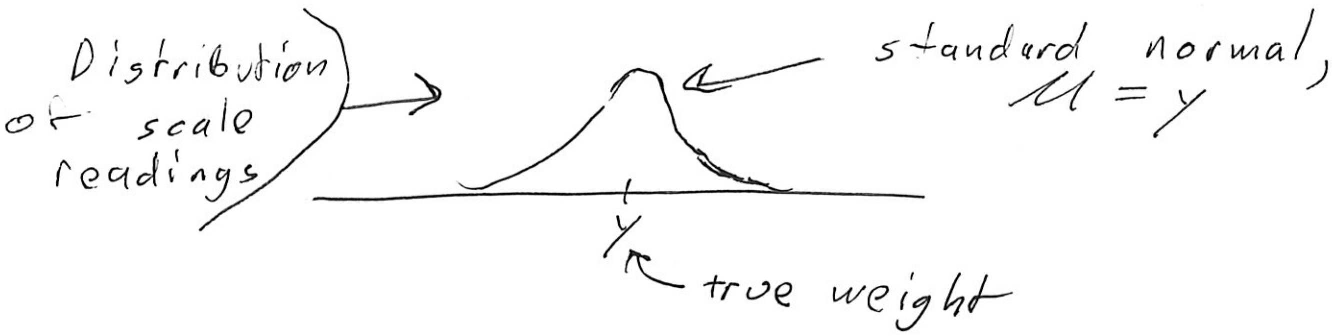
[squares completed in Stigler, chapter 4]

$$A = \underbrace{\frac{\tau^2\mu + \sigma^2 x}{\tau^2 + \sigma^2}}_{\text{Weighted Average of } \mu \text{ and } X} \qquad B^2 = \underbrace{\frac{\tau^2\sigma^2}{\tau^2 + \sigma^2}}_{\text{a posteori uncertainty}}$$

Example:

Measure a weight with an imperfect scale:

Scale makes errors with std. deviation 1 Kg., normally distributed:



X = recorded weight

Y = true weight

$f(x/y) \sim \mathcal{N}(y, 1)$

$f_Y(y) \longrightarrow ??$

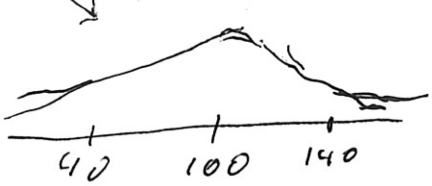
Say $\mathcal{N}(\mu, \sigma^2)$

$\mu = 100$ kg

$\sigma^2 = (10)^2 = 100$

[Why? Maybe have rough idea, say from number of people needed to lift it]

$f_Y(y)$



(A priori: $\left.\begin{array}{l} P(90 < Y \leq 110) \\ P(|Y-100| \leq 10) \end{array}\right\} \approx \frac{2}{3}$

Given "data" x, want $f(y/x)$.

$$f(y/x) \propto f_Y(y)\, f(x/y)$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-y)^2}$$

$$\propto e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2} - \frac{1}{2}(x-y)^2}$$

messy completion of squares not shown

$$= e^{-\frac{1}{2\sigma^2}\left[(y-\mu)^2 + \sigma^2(x-y)^2\right]}$$

$$\propto e^{-\frac{1}{2}\frac{(y-A)^2}{B}} \longleftarrow$$

functional form, aka "Business Part"

$$A = \frac{x\sigma^2}{\sigma^2+1} + \frac{\mu \cdot 1}{\sigma^2+1}$$

weighted average of μ and x.

$$B = \frac{\sigma^2}{(\sigma^2+1)}$$

$f(y|x)$ is $N(A, B)$

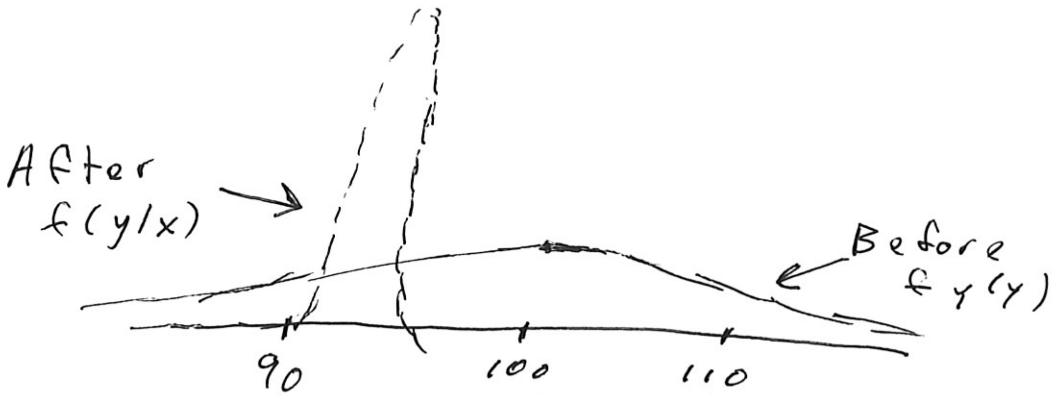
$E(Y/X) = A$ (between X and μ)

If σ^2 small (much prior info)

A near μ

If σ^2 large (little prior info)

A near X



case for $\sigma^2 = (10)^2$, $\mu = 100$
$X = 90$

That is:

$$f_Y(y) \qquad N(100, 10^2)$$

$$f(x/y) \qquad N(y, 1)$$

$$A = \frac{100}{101} \cdot x + \frac{1}{101} \cdot \mu = 90.9$$

$$B = \frac{100}{101}$$

$$f(y/x) \qquad N\left(90.9, \frac{100}{101}\right)$$

Bayes's Theorem Processes Information!

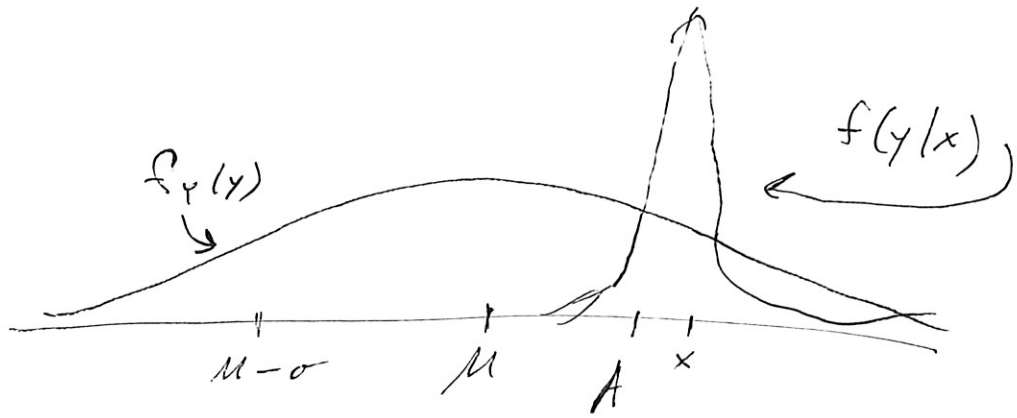
In summary, for the normal dist, we have

$$f(y) \quad \mathcal{N}(\mu, \sigma^2)$$
$$f(x|y) \quad \mathcal{N}(y, 1)$$
$$f(y|x) \quad \mathcal{N}(A, B)$$

$$A = x \cdot \frac{\sigma^2}{\sigma^2+1} + \mu \cdot \frac{1}{\sigma^2+1}$$

$$B = \frac{\sigma^2}{\sigma^2+1}$$

(so if $\lambda = \frac{\sigma^2}{\sigma^2+1}$, $A = x \cdot \lambda + \mu(1-\lambda)$)



Intro to Maximum

Likelihood

Today we'll go further in considering Statistical Inference.

Recall that we would like to know the "state of nature" θ. More exactly θ is a parameter that represents such a state.

We will learn about θ by considering X $[= (x_1, \ldots, x_n)]$, the data.

We need a model to describe the relation between θ and X.

Specifically, X, given θ, is a random variable with distribution $p(x|\theta)$ or $f(x|\theta)$.

Ex: θ = fraction of votes
X = # of 100 sampled

$$p(x|\theta) = \binom{100}{x} \theta^x (1-\theta)^{100-x}$$

$x = 1, 2, \ldots, 100$

Ex: θ = true weight
X = what scale says
$f(x|\theta)$ $N(\theta, 1)$

Ideal Goal: Find $f(\theta/x)$.
ie: After we have data ("given data"), we want to know the probability of various values of θ.

So far, we've used Bayes's Theorem:

$$f(\theta/x) \propto \underbrace{f(\theta)}_{\text{prior}} f(x/\theta)$$

Gives what we want. But:

it requires $f(\theta)$.

$f(\theta)$ is controversial –
How to get it?
what does it mean?
Subjective bias – disagreements

OK. How about a more limited goal?
We won't use $f(\theta)$.
Instead, we will work <u>only</u> with $f(x/\theta)$. Then we'll

<u>Estimate a Point, not a Distribution</u>

We treat θ as fixed (a 'given') X as random. We want an estimate of $\hat{\theta} = f(x)$ that is likely to be close to θ.

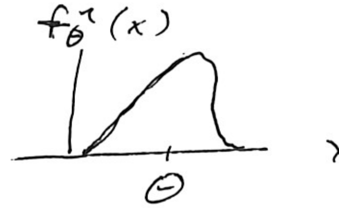
$\hat{\Theta}$ depends on X

$\hat{\theta}$ is short for $\hat{\Theta}(x)$ (or $\hat{\Theta}(x_1, \dots x_n)$)

X is a random variable, so $\hat{\theta}$ is a random variable

---

What does "$\hat{\theta}$ likely to be near θ" mean? From our "given θ" perspective, $\hat{\theta}$ has a distribution $f_{\hat{\theta}}(x)$ or $f_{\hat{\theta}}(x|\theta)$:



We want this distribution to be concentrated near and/or centered at θ.

Defs: $\hat{\theta}$ is <u>unbiased</u> if $E(\hat{\theta}) = \theta$, whatever θ is (ie $\int_{-\infty}^{\infty} x \, f_{\hat{\theta}}(x|\theta)\,dx = \theta$ for all θ).

<u>Bias</u> $= E(\hat{\theta}) - \theta$

<u>Mean Error</u> $= E(|\hat{\theta} - \theta|)$

<u>Mean Square Error</u> $= E[(\hat{\theta} - \theta)^2]$ ("MSE")

It turns out the MSE has a particularly clear interpretation:

$$
\begin{aligned}
MSE_{\hat\theta}(\hat\theta) &= E(\hat\theta - \theta)^2 \\
&= E[(\hat\theta - E(\hat\theta)) + (E(\hat\theta) - \theta)]^2 \\
&= E\Big[(\hat\theta - E(\hat\theta))^2 + 2(\hat\theta - E(\theta))(E(\theta) - \theta) + \left(E(\hat\theta) - \theta\right)^2\Big] \\
&= \underbrace{E[\hat\theta - E(\hat\theta)]^2}_{Var(\hat\theta|\theta)} + 2(E(\hat\theta) - \theta)\underbrace{E(\hat\theta - E(\hat\theta))}_{= E(\hat\theta) - E(\hat\theta) = 0} \\
&\quad + \underbrace{(E(\hat\theta) - \theta)^2}_{(B(\theta))^2}
\end{aligned}
$$

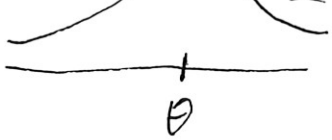
Hence

$$MSE(\hat\theta) = Var(\hat\theta) + (Bias)^2$$

"expected error" — error from spread of data points — error from bias

Tradeoff between bias and variance

Case I — $f_{\hat\theta}$

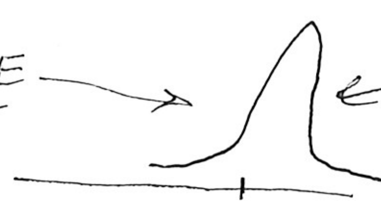


θ

no bias
large variance

SAME MSE

Case II — $f_{\hat\theta}$



θ

bias
small variance

Example: X Binomial(n, θ)

$$\hat{\theta} = \frac{X}{n}. \quad E(\hat{\theta}) = \frac{E(X)}{n} = \frac{n\theta}{n} = \theta$$

unbiased

Example: Same dist, but

$$\hat{\theta}^* = \frac{X+1}{n+2}$$

[This estimator for $\hat{\theta}$ is what we'd get in a Bayesian analysis with $f(\theta)$ uniform on $[0,1]$. Then the posterior dist. would be Beta$(x+1, n-x+1)$ with $E(\theta | X=x) = \frac{x+1}{n+2}$. We are not being Bayesian here, but we can still use the estimator.]

$$E(\hat{\theta}^*) = \frac{E(X)+1}{n+2} = \frac{n\theta+1}{n+2} \neq \theta$$

Biased!