Bayes Rule:

47% of 18-29 year olds use group chat
21% of 30-49 year olds use group chat
7% of >50 year olds use group chat

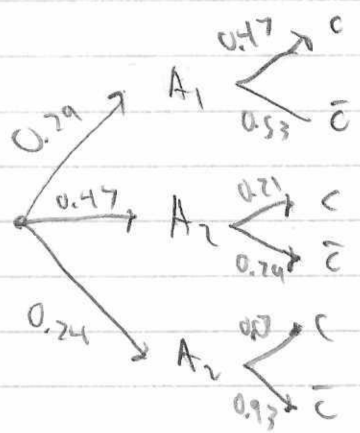
{ should these sum to 100%?
{ No! Each is Pr(group chat | age group)

- If we want the proportion of internet users that use group chat, consider:

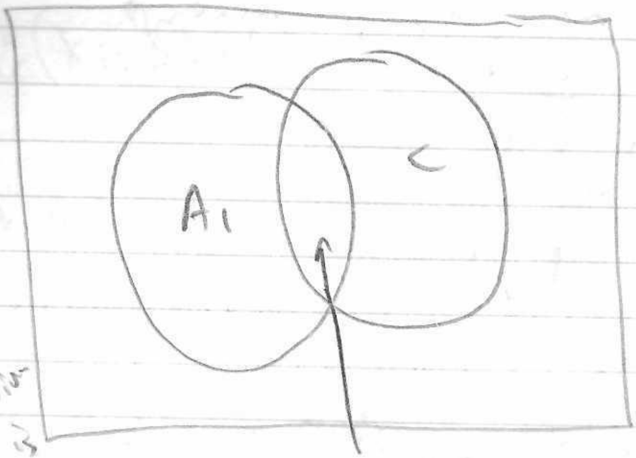
29% of internet users are 18 to 29
47% of internet users are 30 to 49
24% of internet users are >50

C = use group chat
$A_1 = 18-29$
$A_2 = 30-49$
$A_3 = >50$

$$\Pr(C) = \overset{0.47 \cdot 0.29}{\Pr(C|A_1)\Pr(A_1)} + \overset{0.21 \cdot 0.47}{\Pr(C|A_2)\Pr(A_2)} + \overset{0.07 \cdot 0.24}{\Pr(C|A_3)\Pr(A_3)}$$



• But what if we want $Pr(A_1 | C)$



$= Pr(A_1 \text{ and } C)$

what proportion of C-circle is also A_1

$$Pr(A_1 | C) = \frac{Pr(A_1 \text{ and } C)}{Pr(C)}$$

$$\Rightarrow \quad Pr(A_1 \text{ and } C) = Pr(A_1 | C)\, Pr(C)$$

$$Pr(A_1 \text{ and } C) = Pr(C | A_1)\, Pr(C)$$

$$Pr(A_1 | C)\, Pr(C) = Pr(C | A_1)\, Pr(A_1)$$

$$Pr(A_1 | C) = \frac{Pr(C | A_1)\, Pr(A_1)}{Pr(C)}$$

But we just calculated $Pr(C)$

$$= \frac{Pr(C | A_1)\, Pr(A_1)}{Pr(C|A_1) Pr(A_1) + Pr(C|A_2) Pr(A_2) + Pr(C|A_3) Pr(A_3)}$$

Genotype likelihoods

- Often, not given genotypes, but genotype likelihoods

- Let D_i = data for individual i

  X_i = genotype for individual i
  $\in \{0, 1, 2, \ldots, K\}$

  K = Ploidy $K = 2$ for diploids (humans)
  $= 4$ for tetraploids
  $= 6$ for hexaploids

$K = 2 \Rightarrow$ aa aA AA
$K = 4 \Rightarrow$ aaaa aaaA aaAA aAAA AAAA

- Genotype likelihoods

  $\Pr(D_i \mid X_i = k)$ for $k = 0, 1, \ldots, K$

  ↑ don't observe X_i, only observe this

  $\Pr(D_i \mid X_i = 0)$ $\Pr(D_i \mid X_i = 1)$ $\Pr(D_i \mid X_i = 2)$
  0.0001 0.02 0.000004

↑ what is most likely genotype? two!

- But we want $Pr(X_i = k \mid D_i)$

$$Pr(X_i = k \mid D_i) = \frac{Pr(D_i \mid X_i = k)\ Pr(X_i = k)}{Pr(D_i)}$$ genotype posterior

$Pr(D_i \mid X_i = k)$ = genotype likelihood

$Pr(X_i = k)$ = Proportion of individuals w/ genotype k

$Pr(D_i) = Pr(D_i \mid X_i = 0)\, Pr(X_i = 0) + Pr(D_i \mid X_i = 1)\, Pr(X_i = 1) + \dots + Pr(D_i \mid X_i = k)$

- Need to estimate $Pr(X_i = k) =: \pi_k$

$$Pr(D_1, \dots, D_n \mid \pi_k) = \prod_{i=1}^{n} Pr(D_i \mid \pi)$$

$$= \prod_{i=1}^{n} \left(\sum_{k=0}^{k} Pr(D_i \mid X_i = k)\ \pi_k \right)$$

↑ Maximize this over π to get $\hat{\pi}$

then insert that into posterior formula

Exercise: 1% of individuals have a disease

We develop a test s.t. it will detect a disease w.p. 0.99 if an individual has it. But if the individual does not have it, there is a 3% chance of it incorrectly saying they have the disease. You test positive. What is the probability you have the disease?

$$Pr(+ \mid D) = 0.99$$
$$Pr(+ \mid \bar{D}) = 0.97$$
$$Pr(D) = 0.01$$

$$Pr(D \mid +) = \frac{Pr(+ \mid D)\, Pr(D)}{Pr(+ \mid D) Pr(D) + Pr(+ \mid \bar{D})\, Pr(\bar{D})}$$

$$= \frac{0.99 \cdot 0.01}{0.99 \cdot 0.01 + 0.03 \cdot 0.99} \approx 0.25$$

Bayesian Inference

- In frequentist inference, have $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$
θ is some parameter of interest
Estimate θ by maximizing $\sum_{i=1}^{n} \log f(x_i|\theta)$ to get $\hat{\theta}$
Asymptotic sampling distribution $\hat{\theta} \sim N(\theta, [I(\theta)]^{-1})$

fisher information

I θ is fixed (not random)
I p-values and confidence intervals interpreted in terms of sampling distributions.

- Bayesian inference

$X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$
θ is some parameter of interest
θ is random (well, we specify our degree of belief in the value of θ by a probability distribution)

$\pi(\theta)$ = "prior" belief about where θ lies

$$f(\theta|x) = \frac{\pi(\theta) \prod_{i=1}^{n} f(x_i|\theta)}{\prod_{i=1}^{n} f(x_i)}$$ Bayes Rule

$$= \frac{f(x|\theta)\,\pi(\theta)}{f(x)}$$

= "posterior" belief about where θ lies after we see data.

Ex.) I believe human height has an average of about 5'7 = 67"
But I'm not sure, could be 5'4 to 5'10
64" 70"

Not range of heights, range of where I think mean height is.

let $2\sigma = 3" \Rightarrow \sigma = 1.5"$

Prior Belief; $\theta \sim N(67, 1.5^2)$

prior belief in mean height

I see data
$x_1 = 63$
$x_2 = 65$
$x_3 = 66$
$x_4 = 65$
$x_5 = 70$

$x_i \overset{iid}{\sim} N(\theta, \sigma^2)$

What is my posterior belief?

Suppose know σ^2

$$\theta \sim N(\mu, \tau^2)$$

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

$$f(x_1, \dots, x_n \mid \theta) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \theta)^2\right\}$$

$$f(\theta) = (2\pi\tau)^{2} \exp\left\{-\frac{1}{2\tau^2}(\theta - \mu)^2\right\}$$

$$f(\theta \mid x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n \mid \theta) \; f(\theta)}{f(x_1, \dots, x_n)}$$

$\leftarrow$ does not depend on θ

$$\propto f(x_1, \dots, x_n \mid \theta) \, f(\theta)$$

$$\propto \exp\left\{ \sum_{i=1}^{n} \frac{-1}{2\sigma^2}(x_i - \theta)^2 + \frac{-1}{2\tau^2}(\theta - \mu)^2 \right\}$$

$$= \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i^2 - 2x_i\theta + \theta^2) - \frac{1}{2\tau^2}(\mu^2 - 2\theta\mu + \theta^2) \right\}$$

$$\propto \exp\left\{ \theta \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{n}{2\sigma^2}\theta^2 + \frac{\mu}{\tau^2}\theta - \frac{1}{2\tau^2}\theta^2 \right\}$$

$$= \exp\left\{ \left(\frac{n\bar{x}}{\sigma^2} + \frac{\mu}{\tau^2}\right)\theta - \frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)\theta^2 \right\}$$

$$\theta \mid x \sim N(a, b) \Rightarrow f(\theta) \propto \exp\left\{-\frac{1}{2b^2}(\theta - a)^2\right\}$$

$$\propto \exp\left\{-\frac{1}{2b^2}\theta^2 + \frac{a}{b^2}\theta\right\}$$

$$-\frac{1}{2b^2} = -\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right) \Rightarrow b^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}$$

$$\frac{a}{b^2} = \left(\frac{n\bar{x}}{\sigma^2} + \frac{\mu}{\tau^2}\right) \Rightarrow a = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{n\bar{x}}{\sigma^2} + \frac{\mu}{\tau^2}\right)$$

$$a = \frac{\cancel{\sigma^2\tau^2}}{n\tau^2 + \sigma^2} \; \frac{n\tau^2\bar{x} + \sigma^2\mu}{\cancel{\sigma^2\tau^2}} = \frac{n\tau^2\bar{x} + \sigma^2\mu}{n\tau^2 + \sigma^2}$$

$$= \frac{\tau^2}{\tau^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}\mu$$

- Prior variance large $\Rightarrow$ don't have a good idea about θ so most of weight is on $\bar{X}$

- Sample size large $\Rightarrow$ data override prior information, so lots of weight on $\bar{X}$

- Exercise: Flip a coin n times.

$X \sim Bin(n, p) \qquad f(x|p) \propto p^x (1-p)^{n-x}$

want to estimate p

(1) Do this by MLE $\qquad \max_p x \log(p) + (n-x)\log(1-p)$

(2) use prior $p \sim Beta(\alpha, \beta)$

$f(p) \propto p^{\alpha}(1-p)^{\beta} \qquad$ Note: mean of β is $\dfrac{\alpha}{\alpha+\beta}$

$f(p|x)$?

- When you have posterior distribution, all parameter summaries come from that distribution

Point estimate $E[\theta|x]$

Credible interval: l, u s.t. $\Pr(l \leq \theta \leq u) = 0.95$

θ is random here

In CI l and u are random