Statistical Inference

Lecture 4a

ANU - RSFAS

Last Updated: Wed Mar 22 11:05:26 2017

Invariance Property of MLEs

Theorem: If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta}) = \widehat{\tau(\theta)}$.

- If the mapping of $\theta \rightarrow \tau(\theta)$ is one-to-one (for each value of θ there is a unique value of $\tau(\theta)$, and vice versa), then everything is straight-forward. We simply note:

$$\eta = \tau(\theta) \rightarrow \tau^{-1}(\eta) = \theta$$

Invariance Property of MLEs

- Define our likelihood based on the reparameterization ($\theta = \tau^{-1}(\eta)$):

$$L^*(\eta|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\tau^{-1}(\eta)) = L(\tau^{-1}(\eta)|\boldsymbol{x}) = L(\theta|\boldsymbol{x})$$

- We find the supremumum of likelihood

$$\underset{\eta}{sup} L^*(\eta|\boldsymbol{x}) = \underset{\eta}{sup} L(\tau^{-1}(\eta)|\boldsymbol{x}) = \underset{\theta}{sup} L(\theta|\boldsymbol{x})$$

to see that the maximum of $L^*(\eta|\boldsymbol{x})$ is when $\eta = \tau(\theta) = \tau(\hat{\theta})$.

Invariance Property of MLEs

- However, many functions of interest are not one-to-one: $\theta \to \theta^2$.
- We proceed by defining the induced likelihood function of L^* for $\tau(\theta)$

$$L^*(\eta|\boldsymbol{x}) = \sup_{\theta:\tau(\theta)=\eta} L(\theta|\boldsymbol{x})$$

- The value $\hat{\eta}$ that maximizes $L^*(\eta|\boldsymbol{x})$ will be called the MLE of η.

Invariance Property of MLEs

Proof: Let $\hat{\eta}$ denote the value that maximizes $L^*(\eta|\boldsymbol{x})$. Let's show for all values of η that

$$L^*\left(\hat{\eta} = \tau(\hat{\theta})|\boldsymbol{x}\right) \geq L^*(\eta|\boldsymbol{x})$$

$$\begin{aligned}
L^*(\eta|\boldsymbol{x}) &= \sup_{\theta:\tau(\theta)=\eta} L(\theta|\boldsymbol{x}) \\
&\leq \sup_{\theta} L(\theta|\boldsymbol{x}) \\
&= L(\hat{\theta}|\boldsymbol{x}) \\
&= \sup_{\theta:\tau(\theta)=\tau(\hat{\theta})} L(\theta|\boldsymbol{x}) \\
&= L^*(\tau(\hat{\theta})|\boldsymbol{x})
\end{aligned}$$

Invariance Property of MLEs

Eg. Normal: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{normal}(\mu, \sigma^2)$.

- If we want the MLE of μ^2 it is $\widehat{\mu^2} = (\hat{\mu})^2$.
- If we want the MLE of σ it is $\widehat{\sigma} = \sqrt{\hat{\sigma}^2}$.
- This tends to be helpful in a computational sense as well (we can remove bounds on parameters):

$$\sigma^2 = exp(\theta) \quad -\infty < \theta < \infty$$

MLE Computation: Expectation - Maximization (EM) Algorithm

- Presentation adapted from CB & *Computational Statistics*.
- The EM algorithm is a general algorithm to find MLEs when some of the data are missing (or the problem can be set in a manner that there are missing data).
- Suppose we observe all of the data $\boldsymbol{y} = \{y_1, \ldots, y_n\}$, then all we do to find the MLE is maximize:

$$\ell(\boldsymbol{\theta}|\boldsymbol{y})$$

- Suppose we don't observe all the $\boldsymbol{y}$s then based on the notation by Donald Rubin we have $\boldsymbol{y} = (\boldsymbol{y}_{obs}, \boldsymbol{y}_{miss})$.

$$
\begin{aligned}
f(\boldsymbol{y}|\boldsymbol{\theta}) &= f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{miss}|\boldsymbol{\theta}) \\
&= k(\boldsymbol{y}_{miss}|\boldsymbol{y}_{obs}, \boldsymbol{\theta}) g(\boldsymbol{y}_{obs}|\boldsymbol{\theta})
\end{aligned}
$$

- This leads to: $g(\boldsymbol{y}_{obs}|\boldsymbol{\theta}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})}{k(\boldsymbol{y}_{miss}|\boldsymbol{y}_{obs}, \boldsymbol{\theta})}$

$$
\begin{aligned}
log\,[g(\boldsymbol{y}_{obs}|\boldsymbol{\theta})] &= log\,[f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{miss}|\boldsymbol{\theta})] - log\,[k(\boldsymbol{y}_{miss}|\boldsymbol{y}_{obs}, \boldsymbol{\theta})] \\
\ell_{obs}(\boldsymbol{\theta}|\boldsymbol{y}_{obs}) &= \ell_{comp}(\boldsymbol{\theta}|\boldsymbol{y}_{obs}, \boldsymbol{y}_{miss}) - log\,[k(\boldsymbol{y}_{miss}|\boldsymbol{y}_{obs}, \boldsymbol{\theta})]
\end{aligned}
$$

- As $\boldsymbol{y}_{miss}$ is missing, we replace the right side of the equation with its expectation:

$$\begin{aligned} \ell_{obs}(\boldsymbol{\theta}|\boldsymbol{y}_{obs}) &= E\left\{\ell_{comp}(\boldsymbol{\theta}|\boldsymbol{y}_{obs},\boldsymbol{y}_{miss})\Big|\boldsymbol{\theta}',\boldsymbol{y}_{obs}\right\} \\ &\quad -E\left\{log\left[k(\boldsymbol{y}_{miss}|\boldsymbol{y}_{obs},\boldsymbol{\theta})\right]\Big|\boldsymbol{\theta}',\boldsymbol{y}_{obs}\right\} \end{aligned}$$

- The EM algorithm seeks to maximize $\ell(\boldsymbol{\theta}|\boldsymbol{y}_{obs})$ with respect to $\boldsymbol{\theta}$ through the following process:

1. **E step**: Calculate the expectation of the complete likelihood conditional on the observed data and the current value of $\boldsymbol{\theta}$:

$$\begin{aligned} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right) &= E\left\{\ell_{comp}(\boldsymbol{\theta}|\boldsymbol{y}_{obs},\boldsymbol{y}_{miss})\Big|\boldsymbol{\theta}^{(r)},\boldsymbol{y}_{obs}\right\} \\ &= \int \left[\ell_{comp}(\boldsymbol{\theta}|\boldsymbol{y}_{obs},\boldsymbol{y}_{miss})\right] k(\boldsymbol{y}_{miss}|\boldsymbol{y}_{obs},\boldsymbol{\theta})d\boldsymbol{y}_{miss} \end{aligned}$$

2. **M step**: Maximize $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right)$ with respect to $\boldsymbol{\theta}$. Set $\boldsymbol{\theta}^{(r+1)}$ equal to the maximizer of Q.

3. Return to the E step unless a stopping criterion has been reached.

- I will not present a proof that the EM algorithm maximizes $\ell(\boldsymbol{\theta}|\boldsymbol{y}_{obs})$. If you are interested please see either CB Exercise 7.31 or *Computational Statistics* Section 4.2.1.

EM Example

- Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} f(y|\theta)$.

$$f(y|\theta) = p \ \mathrm{normal}(\mu_0, \sigma_0^2) + (1-p) \ \mathrm{normal}(\mu_1, \sigma_1^2)$$

- For this problem generally we have $\theta = (\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, p)$.
- For our example let's simplify the problem and assume $p = \frac{1}{2}, \sigma_0^2 = \sigma_1^2 = 1$
- We have the following likelihood:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\frac{1}{2} \ \mathrm{normal}(\mu_0, 1) + \frac{1}{2} \ \mathrm{normal}(\mu_1, 1)\right]$$

- Directly optimizing this is hard.

- To make things easier, we can introduce latent (missing) variables $Z_1, \ldots, Z_n$.
  - Where $Z_i = 0$ if Y_i is from $\mathrm{normal}(\mu_0, 1)$.
  - Where $Z_i = 1$ if Y_i is from $\mathrm{normal}(\mu_1, 1)$.
- Why is this easier? If we know what normal distribution each Y comes from then it is easy to determine the MLEs for μ_0 and μ_1.
- We can use the EM algorithm, where $\boldsymbol{y}_{miss} = \boldsymbol{z}$.
- Note: Z_i is a Bernoulli random variable. $P(Z_i = 1) = \frac{1}{2}$.

$$L(\theta)_{comp} = \prod_{i=1}^{n} \mathrm{normal}(y_i|\mu_0, 1)^{1-z_i} \mathrm{normal}(y_i|\mu_1, 1)^{z_i}$$

$$
\begin{aligned}
\ell_{comp}(\theta) &= \sum_{i=1}^{n}(1-z_i)log\left(\frac{1}{\sqrt{2\pi}}exp\left(-\frac{1}{2}(y_i-\mu_0)^2\right)\right)\\
&\quad+\sum_{i=1}^{n}z_i log\left(\frac{1}{\sqrt{2\pi}}exp\left(-\frac{1}{2}(y_i-\mu_1)^2\right)\right)\\
&\quad+constants\\
&= -\frac{1}{2}\sum_{i=1}^{n}(1-z_i)(y_i-\mu_0)^2+-\frac{1}{2}\sum_{i=1}^{n}z_i(y_i-\mu_1)^2+constants
\end{aligned}
$$

1. Let's determine $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right)$. Note that z is linear in the log likelihood! Makes our job much easier!

$$\begin{aligned} E\left[\ell_{comp}(\theta)\right] &= -\frac{1}{2}\sum_{i=1}^{n}(1 - E[Z_i|\boldsymbol{y}_{obs}, \theta^r])(y_i - \mu_0)^2 \\ & \quad -\frac{1}{2}\sum_{i=1}^{n} E[Z_i|\boldsymbol{y}_{obs}, \theta^r](y_i - \mu_1)^2 + constants \end{aligned}$$

- We need to determine: $E[Z_i|\boldsymbol{y}_{obs}, \theta^r]$.
- Note: $E[Z_i|\boldsymbol{y}_{obs}; \theta^r] = Pr(Z_i = 1|\boldsymbol{y}_{obs}; \theta^r)$

- We will use Bayes' rule.

$$
\begin{aligned}
Pr(Z_i = 1|\boldsymbol{y}_{obs};\theta^r) &= \frac{f(\boldsymbol{y}_{obs}|Z_i = 1;\theta^r)Pr(Z_i = 1)}{f(\boldsymbol{y}_{obs}|Z_i = 1;\theta^r)Pr(Z_i = 1) + f(\boldsymbol{y}_{obs}|Z_i = 0;\theta^r)Pr(Z_i = 0)} \\
&= \frac{\text{normal}(y_i;\mu_1, 1)\frac{1}{2}}{\text{normal}(y_i;\mu_1, 1)\frac{1}{2} + \text{normal}(y_i;\mu_0, 1)\frac{1}{2}} \\
&= \frac{\text{normal}(y_i;\mu_1, 1)}{\text{normal}(y_i;\mu_1, 1) + \text{normal}(y_i;\mu_0, 1)} \\
&= p_i
\end{aligned}
$$

- So we have:

$$
Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right) = -\frac{1}{2}\sum_{i=1}^{n}(1 - p_i)(y_i - \mu_0)^2 - \frac{1}{2}\sum_{i=1}^{n} p_i(y_i - \mu_1)^2 + constants
$$

2. Let's maximize $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right)$ with respect to μ_0, μ_1. We find:

$$\hat{\mu}_0^{(r+1)} = \frac{\sum_{i=1}^{n}(1-p_i)y_i}{\sum_{i=1}^{n}(1-p_i)}$$

$$\hat{\mu}_1^{(r+1)} = \frac{\sum_{i=1}^{n} p_i y_i}{\sum_{i=1}^{n} p_i}$$

- Recompute p_i with $\hat{\mu}_0^{(r+1)}$ and $\hat{\mu}_1^{(r+1)}$. Thus iterate between the E and M steps till convergence.

```r
## generate data from a bivariate normal:
set.seed(1001)
n <- 1000
z <- rbinom(n, 1, 1/2)

y <- rep(NA, n)
y[z==0] <- rnorm(length(z[z==0]), -2, 1)
y[z==1] <- rnorm(length(z[z==1]), 2, 1)
```

- Beacuse we generated the data, we know z, so we know the MLEs:

$$\hat{\mu}_0 = \bar{y}|_{z=0}$$

```
mean(y[z==0])
```

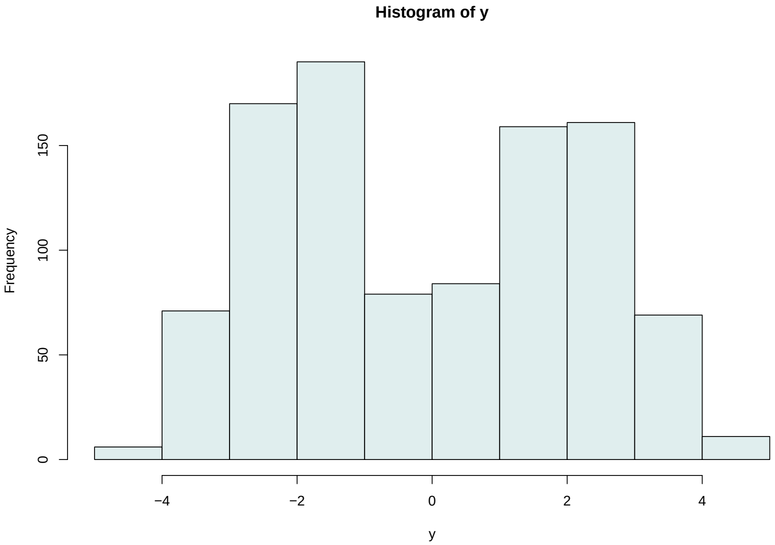
```
## [1] -1.957452
```

$$\hat{\mu}_1 = \bar{y}|_{z=1}$$

```
mean(y[z==1])
```

```
## [1] 1.997553
```

```r
hist(y, col="azure2")
```



E-M

```
## starting values
mu.0 <- -0.5
mu.1 <- 0.5

##
check <- 10
eps <- 1e-10

##
while(check > eps){

  # vector E[z|y] - E step
  rho <- dnorm(y, mu.1, 1)/(dnorm(y, mu.1, 1) + dnorm(y, mu.0, 1))

  # M step
  mu.0.new <- sum((1-rho)*y)/(sum((1-rho)))
  mu.1.new <- sum((rho)*y)/(sum((rho)))

  check <- sum( c(abs(mu.0.new - mu.0), abs(mu.1.new - mu.1)))
  mu.0 <- mu.0.new
  mu.1 <- mu.1.new
}

mu.0.hat <- mu.0
mu.1.hat <- mu.1
```

MLEs based on E-M algorithm

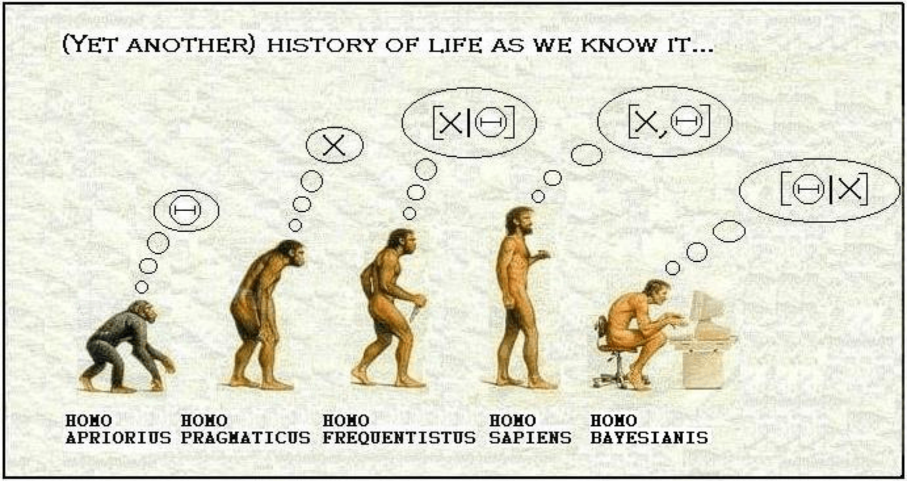
```
mu.0.hat
```

```
## [1] -1.942764
```

```
mu.1.hat
```

```
## [1] 2.007483
```

A Change of Thinking



Some Thoughts on Probability (J. Kadane Text)

- Uncertainty is a fact of life! Let's think about the uncertainty related to the weather tomorrow.
- Let's consider the temperature and whether it rains tomorrow:
  - A_1: Rain and High above 20 degrees C tomorrow
  - A_2: Rain and High at or below 20 degrees C tomorrow
  - A_3: No Rain and High above 20 degrees C tomorrow
  - A_4: No Rain and High at or below 20 degrees C tomorrow.
- We know that one and only one of these will occur tomorrow (exhaustive and disjoint).
- Whatever you are uncertain about, I can ask you to specify a set of disjoint and exhaustive events that describe your categories.

- Consider a ticket that pays:
  - \$1 if event A_1 happens
  - \$0 if A_1 does not happen.
- A buyer of such a ticket pays the seller the amount p.
  - If the event A_1 occurs, the seller pays the buyer \$1.
  - If the event A_1 does not occur, the seller owes the buyer nothing.

- Let's suppose that your price for a \$1 ticket on A_1 is $Pr\{A_1\} = p$ (pronounced 'price of A_1') is 30 cents.
    - If I sell the ticket to you and it rains tomorrow and the temperature is above 20 degrees C, I would have to pay you \$1.

$$\Rightarrow \text{ your total: } -0.30 + 1.00 = 0.70$$

    - If it does not rain or if the temperature does not rise to be above 20 degrees C, I would not pay you anything.

$$\Rightarrow \text{ your total: } -0.30 + 0 = -0.30$$

- Similarly you can name prices for the events A_2, A_3, and A_4.
- Once you set a price, then you are willing to buy or sell tickets at that price.
- It would be foolish for you to specify prices that have the property that I can be assured of making money from you, whatever the weather might be tomorrow (i.e., making you a sure loser).
- Avoiding being a sure loser does not make you a winner, or even likely to be a winner.

- Suppose you were willing to sell a ticket for a negative amount, say $Pr\{A_1\} = -\$0.05$.
- I buy a ticket from you. You give me the ticket and \$0.05.
  - If event A_1 happens:

$$\Rightarrow \text{your total: } (-0.05)\ +\ (-1.00) = -1.05$$

  - If event A_1 does not happen:

$$\Rightarrow \text{your total: } (-0.05)\ +\ 0 = -0.05$$

- No matter what happens, you are a sure loser.
- To avoid this, for every event A:

$$Pr\{A\} \geq 0$$

- Now consider the sure event $S = \{A_1 \ \cup \ A_2 \ \cup \ A_3 \ \cup \ A_4\}$.
- What price should you give to the sure event S?
- If you give a price below \$1, say \$0.75 and I purchase a ticket you are sure to lose \$0.25:

$$\Rightarrow \text{your total: } (0.75) \ + \ (-1.00) = -0.25$$

- If you give a price above \$1, say \$1.25, and I sell you a ticket you are sure to lose \$0.25:

$$\Rightarrow \text{your total: } (-1.25) \ + \ (1.00) = -0.25$$

- To make sure that you are not a sure loser you would like,

$$Pr\{S\} = 1$$

- Suppose you were willing to state the following prices for tickets:

$$\begin{aligned} Pr\{A_1\} &= \$0.20 \\ Pr\{A_2\} &= \$0.25 \\ Pr\{A_1 \ \cup \ A_2\} &= \$0.40 \end{aligned}$$

- I will sell you a ticket for A_1 and a ticket for A_2.
- And buy a ticket for $[A_1 \ \cup \ A_2]$.
- Recall:
  - A_1: Rain and High above 20 degrees C tomorrow
  - A_2: Rain and High at or below 20 degrees C tomorrow

- Suppose it does not rain. Then none of the tickets have to be settled by payment.
    - You gave me $\$0.20 + \$0.25 = \$0.45$.
    - I gave you \$0.40.
    - You lost \$0.05.
- Suppose it does rain. Then either A_1 or A_2 occurred (only one).
    - I owe you \$1.
    - You owe me \$1.
    - In addition I still have the \$0.05 that I gained from the sale and purchase of the tickets.
- Either way, you lose \$0.05.

- The problem that you named too low a price for the ticket on the union $[A_1 \ \cup \ A_2]$.
- Any price less than \$0.45 leads to sure loss for you.

- Now suppose you raise the price on the union: $Pr\{A_1 \cup A_2\} = \$0.60$.
- I will now sell you the $[A_1 \cup A_2]$.
- I buy a ticket for A_1 and A_2.
- Suppose it does not rain. Then none of the tickets have to be settled by payment.
  - You gave me \$0.60.
  - I gave you $\$0.20 + \$0.25 = \$0.45$.
  - You lost \$0.15.
- Suppose it does rain. Then either A_1 or A_2 occurred (only one).
  - I owe you \$1.
  - You owe me \$1.
  - In addition I still have the \$0.15 that I gained from the sale and purchase of the tickets.

- Hence if your price for the ticket on the union of the two events is too low or too high, you can be made a sure loser. Unless,

$$Pr\{A_1\} + Pr\{A_2\} = Pr\{A_1 \ \cup A_2\}$$

- Good news: If your prices satisfy those conditions, you cannot be made a sure loser.
- Prices satisfying these equations are said to be coherent.
- These prices satisfy the Kolmogorov's probability axioms. So based on decisions in the face of uncertainty we end up with probability!
- But note that the equations don't tell you exactly what prices to assign, as long as the equations are satisfied.
- We can each have different beliefs about the outcome of events and they can all be coherent.
- Personalistic (subjective view) of probability.
- From a set of 7 axioms on choice (extending the work of John von Neumann and Oskar Morgenstern), Leonard Savage showed that you can derive utility, probability, conditional probability, all leading to the fact that in the face of uncertainty you should minimize posterior expected loss (a bit more to come on this).

Bayesian Inference

- Bayesian inference is the application of Bayes' rule to the problem of 'guessing' an unknown quantity(ies).
- The key distinction between Bayesian and sampling theory statistics is the issue of what is to be regarded as random and what is to be regarded as fixed.
    - To a Bayesian, parameters are random and data, once observed, are fixed.
    - To a sampling theorist (classical statistician), data are random even after being observed (how inference is done), but parameters are fixed.

A Very Stylized Example

- Let's consider an example that was discussed in *A First Course in Bayesian Statistics* by Hoff (2009).
- Consider determining the prevalence of an infectious disease in a particular city - say Canberra.
- Our interest lies in the fraction of infected individuals in Canberra (θ).
- We plan on taking a random sample of 20 individuals and determine how many of them are infected (y).

$$y|\theta \sim \mathrm{binomial}(20, \theta)$$

- We conduct our sample and find that none of the individuals are infected ($y = 0$).
- What is our inference (best guess) for θ?

Inferential Approach 1 - MLE

- We can use *Maximum Likelihood Estimation*:

$$\begin{aligned}\hat{\theta} &= \mathit{argmax}_{\theta} \;\; L(\theta|y) \\ &= \mathit{argmax}_{\theta} \;\; \left(\begin{array}{c} N \\ y \end{array} \right) \theta^{y}(1-\theta)^{N-y} \\ &= y/N = 0/20 = 0.\end{aligned}$$

Inferential Approach 2 - Bayes

- Suppose in other studies of similar cities the infection rate ranges from about 0.05 to 0.20 with an average prevalence of 0.09.
- A *beta* distribution with parameters $a = 2, b = 20$ has a mean value of 0.091.

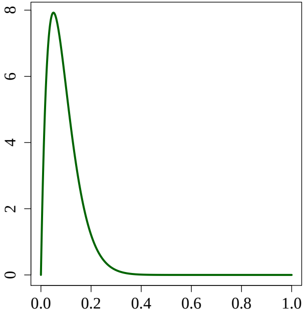


Figure 1: Proportion infected in the population.

Inferential Approach 2 - Bayes

- Thus far the information we have is:

$$p(y|\theta) = \mathrm{binomial}(20, \theta) \quad \& \quad p(\theta) = \mathrm{beta}(2, 20)$$

- Now instead of maximizing we will use Bayes' rule to infer θ:

$$p(\theta|y) = \frac{p(y \text{ and } \theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int_0^1 p(y|\theta)p(\theta)d\theta}$$

- If we do a little math $\Rightarrow \theta|y \sim \mathrm{beta}(2, 40)$.

Inferential Approach 2 - Bayes

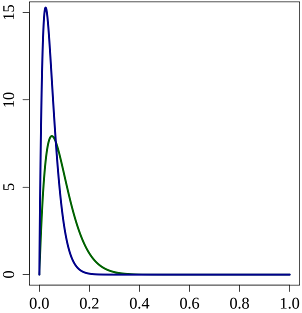


Figure 2: Proportion infected in the population (posterior in blue [$p(\theta|y)$], prior in green [$p(\theta)$]

Inferential Approach 2 - Bayes

- What is our estimate? The Bayesian approach provided a full distribution! We could use the mean or median as a point estimate:

$$E(\theta|y) = 2/42 \approx 0.048$$

- Or provide *95% credible interval* for $\theta|y$:

$$Pr(0.006 \leq \ [\theta|y] \ \leq 0.129) \approx 0.95$$

Inferential Approach 1 - MLE

- Back to the Maximum Likelihood estimation.
- Confidence intervals for ML estimators:

$$\hat{\theta} \pm 1.96 \times \mathrm{S.E.}(\hat{\theta})$$

How are these intervals interpreted?

- In our case:

$$\hat{\theta} \pm 1.96 \times \sqrt{\hat{\theta}(1-\hat{\theta})/N} = [0,0]$$

- Of course there are frequentist approaches to deal with estimation of small proportions!

Similar Example - Han Solo and Bayesian Priors

- For a fun similar example:

https://www.statslife.org.uk/the-statistics-dictionary/2148-han-solo-and-bayesian-priors

Why consider Bayesian inference?

- Bayesian methods provide the user with the ability to formally incorporate prior information (Carlin and Louis, 2009).
- If $p(y|\theta)$ and $p(\theta)$ represent a rational person's beliefs then Bayes' rule is an optimal method for updating those beliefs.
- If $p(y|\theta)$ and $p(\theta)$ approximately represents beliefs then $p(\theta|y)$ is also approximate and may be useful (Hoff 2009).

"All models are wrong, but some are useful" - Box and Draper.

Why consider Bayesian inference?

- Inferences are conditional on the actual data (Carlin and Louis, 2009).
- Bayesian answers are more easily interpretable by non-statisticians (Carlin and Louis, 2009).
- All Bayesian analyses follow *directly* from the posterior; no separate theories of estimation, testing, multiple comparisons, etc. are needed (Carlin and Louis, 2009).
- In many complicated statistical problems there are no obvious non-Bayesian methods of estimation or inference (Hoff 2009).

Bayesian Inference

- Bayesian Inference is simply the application of Bayes' Rule to infer about parameters:

$$\begin{aligned}
\pi(\theta|\boldsymbol{x}) &= \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{\int_{\Theta} f(\boldsymbol{x}|\theta)\pi(\theta)d\theta} \\
&= \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{m(\boldsymbol{x})} \\
&\propto f(\boldsymbol{x}|\theta)\pi(\theta)
\end{aligned}$$

- It is important to remember though that in a Bayesian framework θ is random. In the frequentist framework it is fixed!

Bayesian Inference

- $\pi(\theta|\boldsymbol{x})$ is the posterior distribution for θ.
- $f(\boldsymbol{x}|\theta)$ is the joint sampling distribution for $\boldsymbol{X}$ or the likelihood for θ.
- $\pi(\theta)$ is the prior distribution for θ.
- $m(\boldsymbol{x})$ is the marginal distribution for $\boldsymbol{X}$.

- Bayesian approach:
  - We start with a prior belief about a situation (represented through a probability distribution parameterized by θ).
  - We observe data $\boldsymbol{x}$.
  - Given the data $\boldsymbol{x}$, we update our beliefs about θ via Bayes' rule and obtain the posterior distribution.

Bayesian Inference

- Many times you will just see this notation:

$$
\begin{aligned}
p(\theta|x_1,\ldots,x_n) &= \frac{p(x_1,\ldots,x_n|\theta)p(\theta)}{\int_\Theta p(x_1,\ldots,x_n|\theta)p(\theta)d\theta} \\
&= \frac{p(x_1,\ldots,x_n|\theta)p(\theta)}{p(x_1,\ldots,x_n)} \\
&= \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})} \\
&\propto p(\boldsymbol{x}|\theta)p(\theta)
\end{aligned}
$$

Bayesian Inference

- As $\pi(\theta|\boldsymbol{x})$ is a probability distribution function, we can consider a number of summaries for θ after we observe the data:

- Posterior mode: the value of θ which maximizes the posterior distribution
- Posterior median
- Posterior mean

Bayesian Inference

- For reasons of mathematical simplicity (though we shall see there are other good reasons), we shall focus on the posterior mean:

$$\hat{\theta}_B = E\left\{[\theta|\boldsymbol{x}]\right\} = \int_\Theta \theta\pi(\theta|\boldsymbol{x})d\theta$$

- We can also consider functions of θ, $\tau(\theta)$:

$$\widehat{\tau(\theta)}_B = E\left\{[\tau(\theta)|\boldsymbol{x}]\right\} = \int_\Theta \tau(\theta)\pi(\theta|\boldsymbol{x})d\theta$$

Bayesian Inference

Eg. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$. Additionally assume:

$$\pi(\theta) = 1 \quad \forall\ \theta$$

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

Bayesian Inference

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

$$\begin{aligned} m(\boldsymbol{x}) &= \int_0^1 f(\boldsymbol{x}|\theta)\pi(\theta)d\theta \\ &= \int_0^1 \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}(1)d\theta \end{aligned}$$

- Math fact (beta function):

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

Bayesian Inference

$$
\begin{aligned}
m(\boldsymbol{x}) &= \int_0^1 \theta^{(\sum x_i + 1) - 1}(1-\theta)^{(n - \sum x_i + 1) - 1} d\theta \\
&= \frac{\Gamma(\sum x_i + 1)\Gamma(n - \sum x_i + 1)}{\Gamma([\sum x_i + 1] + [n - \sum x_i + 1])}
\end{aligned}
$$

Bayesian Inference

- So our posterior distribution for θ is:

$$\begin{aligned}
\pi(\theta|\boldsymbol{x}) &= \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{m(\boldsymbol{x})} \\
&= \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}(1)}{\frac{\Gamma(\sum x_i+1)\Gamma(n-\sum x_i+1)}{\Gamma([\sum x_i+1]+[n-\sum x_i+1])}} \\
&= \frac{\Gamma([\sum x_i+1]+[n-\sum x_i+1])}{\Gamma(\sum x_i+1)\Gamma(n-\sum x_i+1)}\theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \\
&= \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)}\theta^{\alpha^*-1}(1-\theta)^{\beta^*-1}
\end{aligned}$$

$$\alpha^* = \sum x_i + 1$$

$$\beta^* = n - \sum x_i + 1$$

Bayesian Inference

$$[\theta|\boldsymbol{x}] \sim \mathrm{beta}(\alpha^*, \beta^*)$$

- Using the posterior distribution of θ and use the posterior mean as an estimate! It turns out the posterior mean is:

$$\begin{aligned}\hat{\theta}_B &= \frac{\alpha^*}{\alpha^* + \beta^*} \\ &= \frac{\sum_{i=1}^{n} X_i + 1}{n+2}\end{aligned}$$

Bayesian Inference

Eg. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$. Additionally assume:

$$\theta \sim \text{beta}(a, b)$$

We can re-write that mean of the posterior:

$$\begin{aligned}
\hat{\theta}_B &= \frac{a + \sum_{i=1}^n X_i}{n + a + b} \\
&= \frac{n}{n+a+b}\bar{X} + \frac{a+b}{n+a+b}\left(\frac{a}{a+b}\right) \\
&= \frac{n}{n+a+b} \times \text{ data average } + \frac{a+b}{n+a+b} \times \text{ prior expectation} \\
&= \frac{n}{n+a+b} \times \text{ MLE } + \frac{a+b}{n+a+b} \times \text{ prior expectation}
\end{aligned}$$

Bayesian Inference

Definition: Let $\mathcal{F}$ denote the class of pdfs or pmfs $f(x|\theta)$. A class $\mathcal{P}$ of prior distributions is a conjugate family for $\mathcal{F}$ if the posterior distribution is in the class $\mathcal{P}$ for all $f \in \mathcal{F}$, all priors in $\mathcal{P}$, and all $x \in \mathcal{X}$.

- In other words, if the prior distribution is in the same family of distributions as the posterior, then it is a conjugate prior distribution.
- For $f(\boldsymbol{x}|\theta)$ examine the kernel with regard to θ and see if your recognize the distribution. This will be the conjugate distribution.

Bayesian Inference

Eg. Let $X_1, \ldots, X_n \sim \mathrm{Poisson}(\lambda)$.

- Determine a conjugate prior distribution and then determine the posterior distribution for λ.

$$\begin{aligned} f(\boldsymbol{x}|\lambda) &= \frac{\lambda^{\sum x_i} e^{-n\lambda}}{\prod x_i!} \\ &\Rightarrow \lambda^{\sum x_i} e^{-n\lambda} \end{aligned}$$

- This is a kernel for a gamma distribution. Thus we have the following conjugate prior:

$$\lambda \sim \mathrm{gamma}(a, b)$$

Bayesian Inference

$$\begin{aligned} p(\lambda|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\lambda)p(\lambda) \\ &\propto \left[\lambda^{\sum x_i} e^{-n\lambda}\right]\left[\lambda^{a-1}e^{-\lambda/b}\right] \\ &= \lambda^{\sum x_i+a-1}e^{-\lambda(n+1/b)} \\ &= \lambda^{a^*-1}e^{-\lambda/b^*} \end{aligned}$$

$$[\lambda|\boldsymbol{x}] \sim \mathrm{gamma}(a^*, b^*)$$

$$a^* = \sum x_i + a$$

$$b^* = \frac{b}{bn+1}$$

Bayesian Estimation

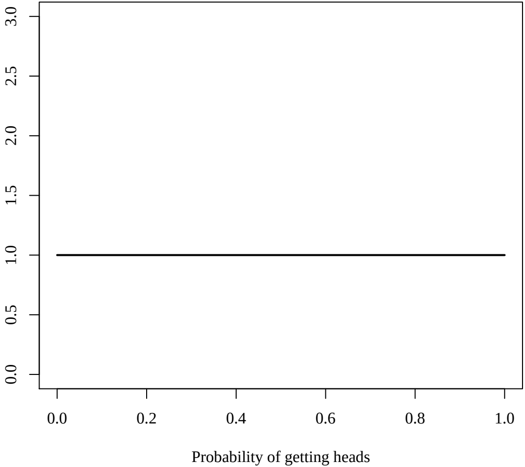
- Are we restricted by conjugate priors? Do they represent my beliefs?
- No! A mixture of conjugate priors is also conjugate.

$$p(\theta) = pf_1(\theta) + (1 - p)f_2(\theta)$$

- Eg. Example by Diaconis and Ylvisaker (1985)
- When a coin is spun on its edge, instead of being thrown in the air, the proportion of *heads* is rarely close to $1/2$, but rather $1/3$ or $2/3$ because of irregularities in the edge that cause the [coin] to favor one side or the other.
- So suppose I have a coin, before I spin it what might my prior beliefs be about the probability of getting a heads?

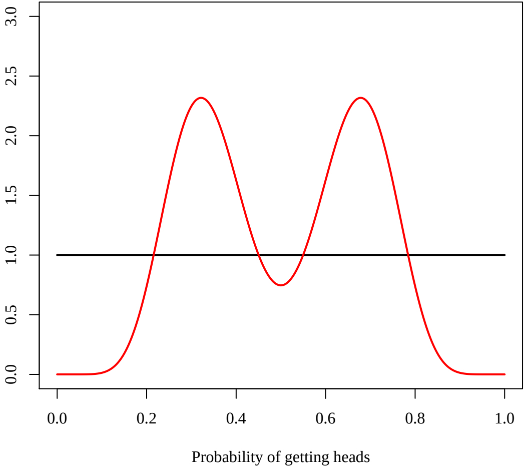
Bayesian Estimation

$$p(\theta) = \mathrm{beta}(a = 1, b = 1)$$



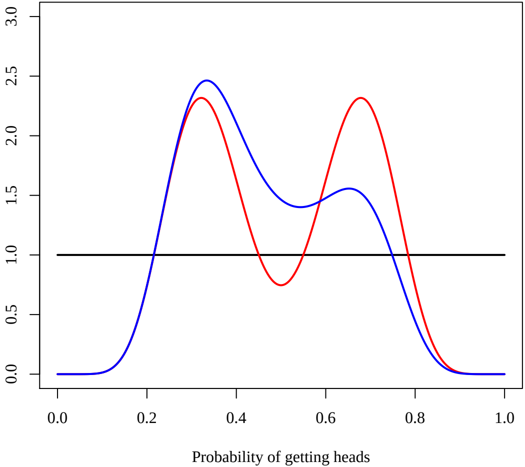
Bayesian Estimation

$$p(\theta) = \frac{1}{2}\mathrm{beta}(a = 10, b = 20) + \frac{1}{2}\mathrm{beta}(a = 20, b = 10)$$



Bayesian Estimation

$$p(\theta) = 0.5\mathrm{beta}(a = 10, b = 20) + 0.2\mathrm{beta}(a = 15, b = 15) + 0.3\mathrm{beta}(a = 20, b = 10)$$



Bayesian Estimation

- Now suppose I flip the coin 10 times and I see:

H T T T H T H T T T

- Now update via Bayes' rule. We get the following posteriors:
    - $p(\theta|\boldsymbol{x}) = \mathrm{beta}(1 + y, n - y + 1)$, i.e., $\mathrm{beta}(4, 8)$;
    - $p(\theta|\boldsymbol{x}) = 0.89\mathrm{beta}(13, 27) + 0.11\mathrm{beta}(23, 17)$; and
    - $p(\theta|\boldsymbol{x}) = 0.77\mathrm{beta}(13, 27) + 0.17\mathrm{beta}(18, 22) + 0.06\mathrm{beta}(23, 17)$.

Bayesian Estimation

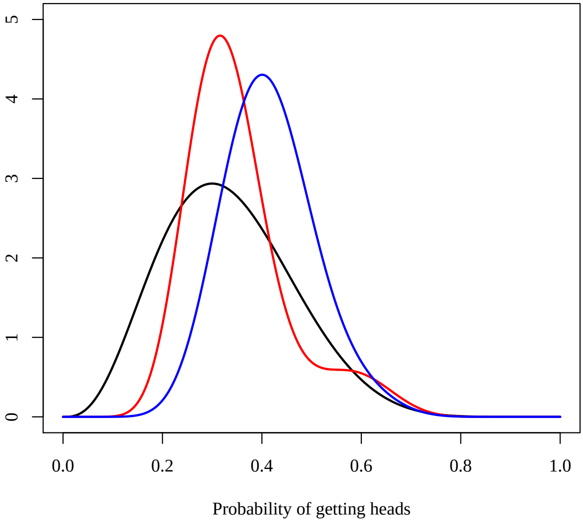


Figure 6: Posterior Beliefs

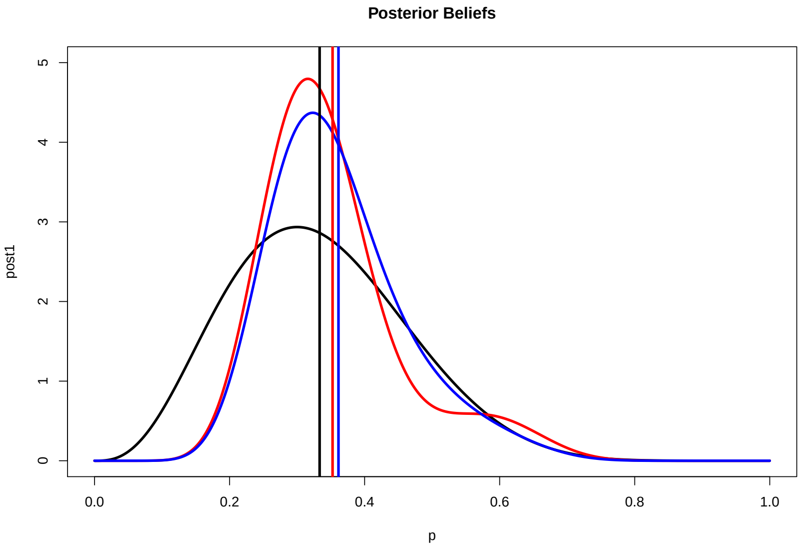
Bayesian Estimation

```r
p <- seq(0,01, by=0.001)
post1 <- dbeta(p, 4, 8)
post2 <-  0.89*dbeta(p, 13, 27) + 0.11*dbeta(p, 23,17)
post3 <-  0.77*dbeta(p, 13, 27) + 0.17*dbeta(p, 18, 22) +
  0.06*dbeta(p, 23, 17)

m.post1 <- 4/(4+8)
m.post2 <- 0.89*( 13/(13+27) ) + 0.11*( 23/(23+17) )
m.post3 <- 0.77*( 13/(13+27) ) + 0.17*( 18/(18+22) ) +
  0.06*( 23/(23+17) )

plot(p, post1, type="l", lwd=3, col="black",
     main="Posterior Beliefs", ylim=c(0,4))
lines(p, post2, col="red", lwd=3)
lines(p, post3, col="blue", lwd=3)
abline( v=c(m.post1, m.post2, m.post3),
        col=c("black", "red", "blue"), lwd=3)
```

Bayesian Estimation



- Posterior means: 0.3333333, 0.3525, 0.36125.