LECTURE OUTLINE
The Big Picture

Professor Leibon

Math 50

Feb. , 2004

Goals

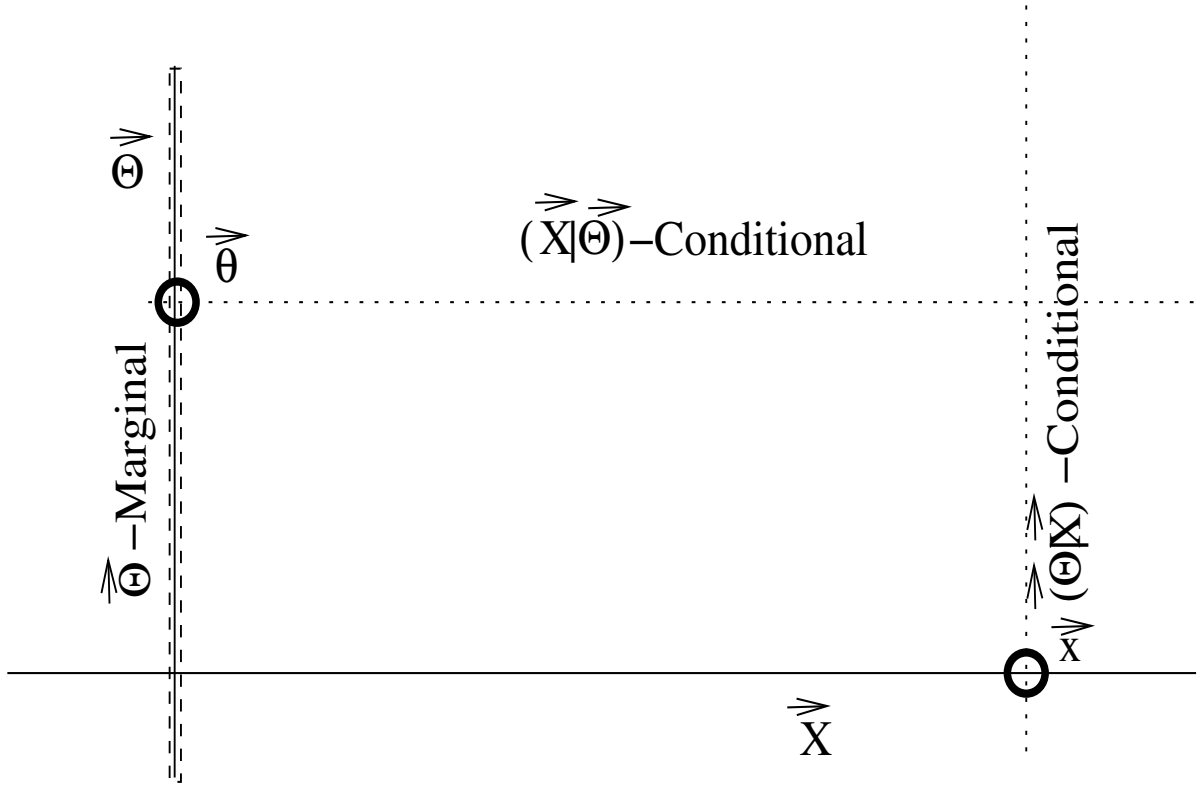
Question: Why have we suffered?

Answer: To develop the Bayesian Framework.

Look at our two examples in this context.

Our Terminology

First we give some notation associated to the below picture in the presence of a joint density function $f_{(\vec{X},\vec{\Theta})}(\vec{x},\vec{\theta})$. Our notation will look like this.



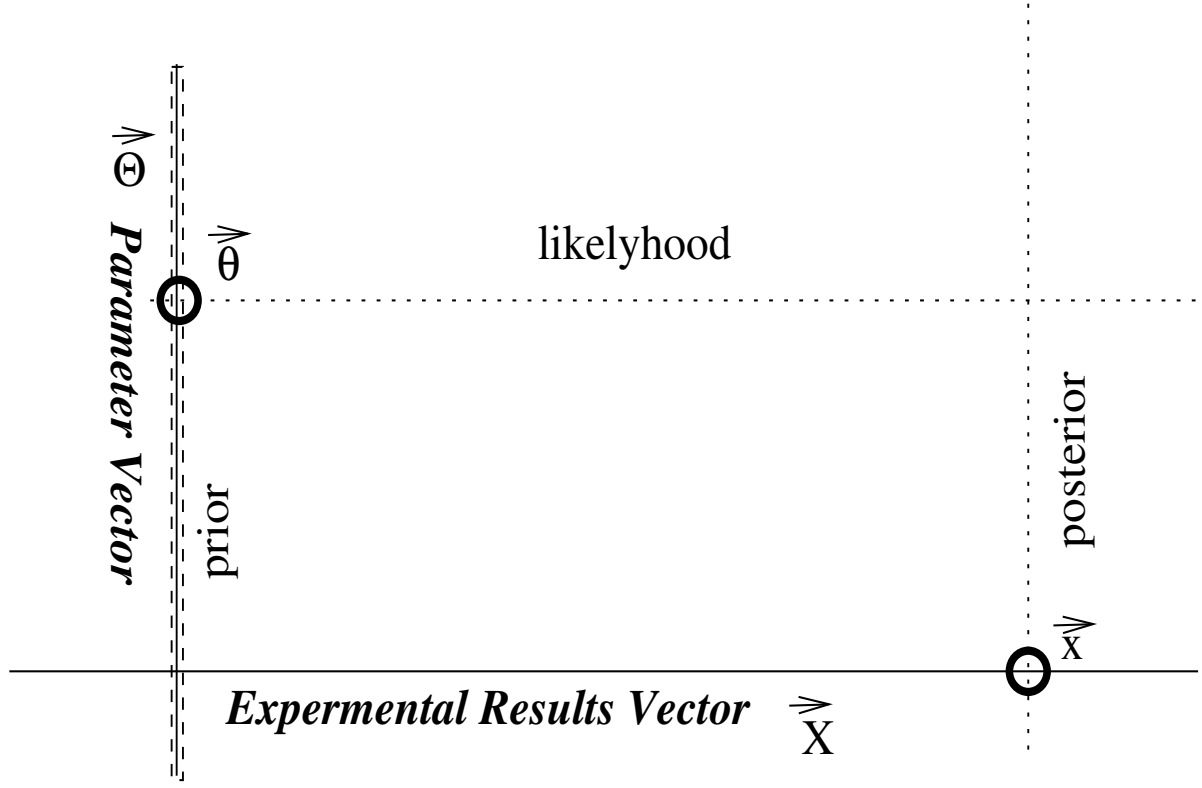
Comparing Terminology

Here is a compare and contrast to help keep track of terminology and notation.

Probability	Statistics	Book	Alternate
$\vec{\theta}$-Values	Parameters	$\vec{\theta}$	$\vec{\theta}$
$\vec{\Theta}$-Marginal	Prior	$f_2(\vec{\theta})$	$f_{\vec{\Theta}}(\vec{\theta})$
$\vec{x}$-Values	Results	$\vec{x}$	$\vec{x}$
$\vec{X}$-Marginal	Nomalizer	$f_1(\vec{x})$	$f_{\vec{X}}(\vec{x})$
$(\vec{\Theta}\|\vec{X})$-Conditional	Posterior	$g(\vec{\theta}\|\vec{x})$	$f_{(\vec{\Theta}\|\vec{X}=\vec{x})}(\vec{\theta})$
$(\vec{X}\|\vec{\Theta})$-Conditional	Likleyhood	$g(\vec{x}\|\vec{\theta})$	$f_{(\vec{X}\|\vec{\Theta}=\vec{\theta})}(\vec{x})$
Joint Density	Joint	$f(\vec{x},\vec{\theta})$	$f_{(\vec{X},\vec{\Theta})}(\vec{x},\vec{\theta})$

A Picture of the Bayesian Terminology

The dotted/dashed lines indicate the set on which the given function is a pdf of pf.



Important Conventions

1. $f_{\vec{Z}}$ can be mixture of the discrete and the continuos, though for our statistical purposes this mixture will always be such that each one dimensional marginal f_{Z_i} is **either** discrete **or** continuous. In particular,

$$\int f_{(\vec{X},\vec{Y})}(\vec{x},\vec{y})d\vec{y}$$

denotes an integral over the continuous $\vec{Y}$ coordinates and a sum over the discrete $\vec{Y}$ coordinates.

2. Let $\chi_A(p)$ be the function which is 1 if $p \in A$ and zero otherwise. This is called an *indicator function*.

The Bayesian View

Facts: We **always** know the likelihood function $f_{(\vec{X}|\vec{\Theta}=\vec{\theta})}(\vec{x})$. If we also believe we can determine the prior $f_{\vec{\Theta}}(\vec{\theta})$ then we are doing we doing *Bayesian statistics*. Bayes' Theorem dictates the posterior density (or probability function) is given by

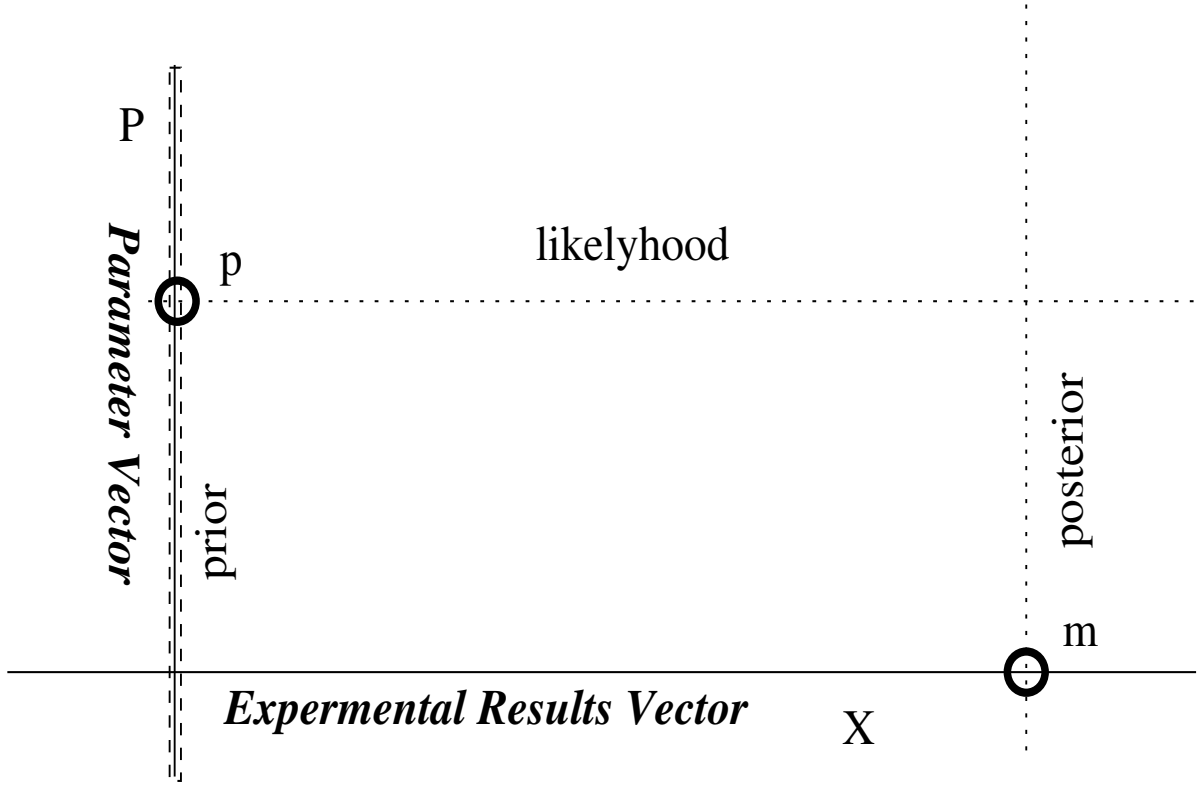
$$f_{(\vec{\Theta}|\vec{X}=\vec{x})}(\vec{\theta}) \propto f_{(\vec{X}|\vec{\Theta}=\vec{\theta})}(\vec{x}) f_{\vec{\Theta}}(\vec{\theta}).$$

Note: We may *normalize* this function via

$$f_{(\vec{\Theta}|\vec{X}=\vec{x})}(\vec{\theta}) = \frac{f_{(\vec{X}|\vec{\Theta}=\vec{\theta})}(\vec{x}) f_{\vec{\Theta}}(\vec{\theta})}{\int f_{(\vec{X}|\vec{\Theta}=\vec{\theta})}(\vec{x}) f_{\vec{\Theta}}(\vec{\theta}) d\vec{\theta}} = \frac{f_{(\vec{X}|\vec{\Theta}=\vec{\theta})}(\vec{x}) f_{\vec{\Theta}}(\vec{\theta})}{f_{\vec{X}}(\vec{x})}.$$

Example 1: Discrete Example

$\vec{\Theta} = P$ discrete with $f_P(p_0) = P_0$ and $f_P(p_1) = (1 - P_0)$.
$\vec{X} = X$ discrete with domain $\{0, \ldots, m\}$ and
$f_{(X|P=p)}(m) = \binom{N}{m} p^m (1-p)^{N-m}$.



Example 1: Discrete Example

Hence

$$f_{(P|X=m)}(p) \propto f_P(p)p^m(1-p)^{N-m}$$

Summing, we find

$$f_{(P|X=m)}(p) = \frac{f_P(p)p^m(1-p)^{N-m}}{(P_0)p_0^m(1-p_0)^{N-m} + (1-P_0)p_1^m(1-p_1)^{N-m}}$$

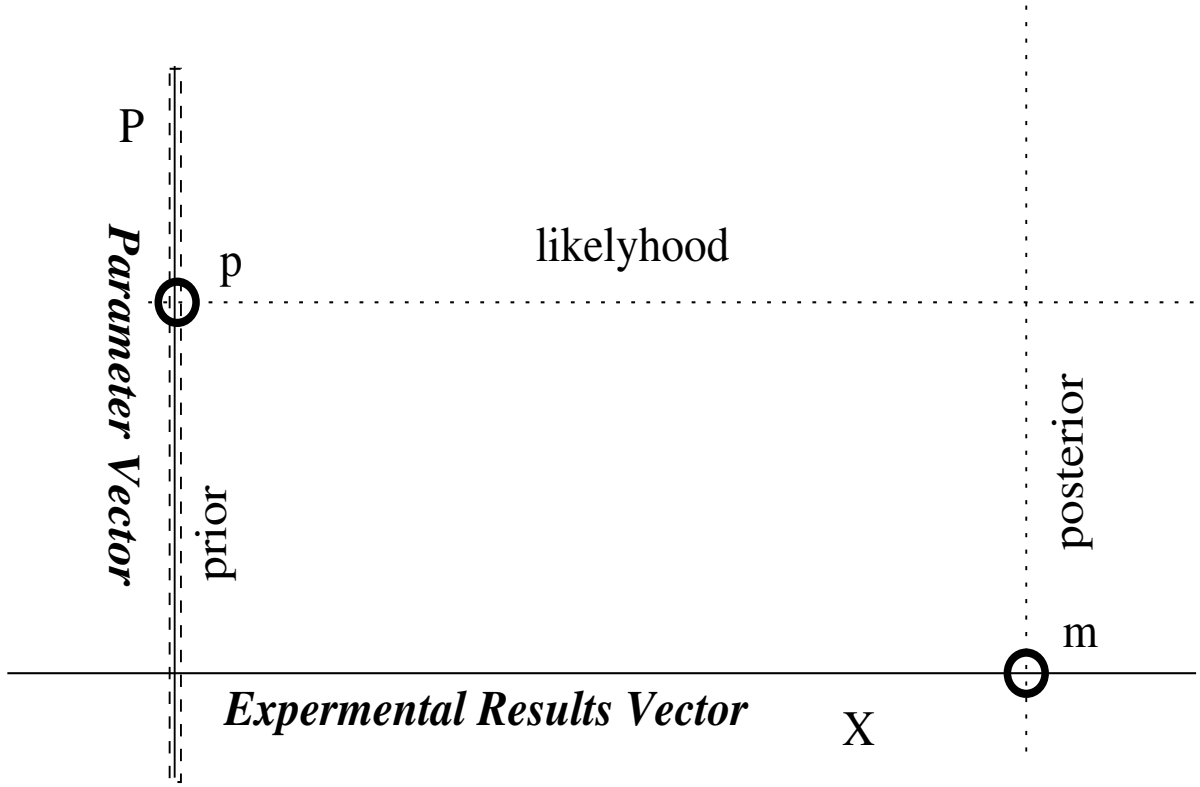
Application: Apply when $P_0 = 0.9$, $p_0 = 0.2$ and $p_1 = 0.7$ and explain the relation to our Madako Experiment. (In particular, find where $Pr(\theta = \theta_1|X = 1) = 0.003$ and recompute this using the above $f_{(P|X=m)}(p)$ function.)

Example 2: "Clinical Trial"

$\vec{\Theta} = P$ continuous with $f_P(p) = \frac{l!k!}{(l+k+1)!} p^k (1-p)^l \chi_{[0,1]}(p)$.

$\vec{X} = X$ discrete with domain $\{0, \ldots, m\}$ and

$f_{(X|P=p)}(m) = \binom{N}{m} p^m (1-p)^{N-m}$.



Example 2: "Clinical Trial"

Hence

$$f_{(P|X=m)}(p) \propto p^{m+k}(1-p)^{N-m+l}\chi_{[0,1]}(p)$$

Integrating, we find

$$f_{(P|X=m)}(p) = \frac{(m+k)!(N-m+l)!}{(N+k+l+1)!}p^{m+k}(1-p)^{N-m+l}\chi_{[0,1]}.$$

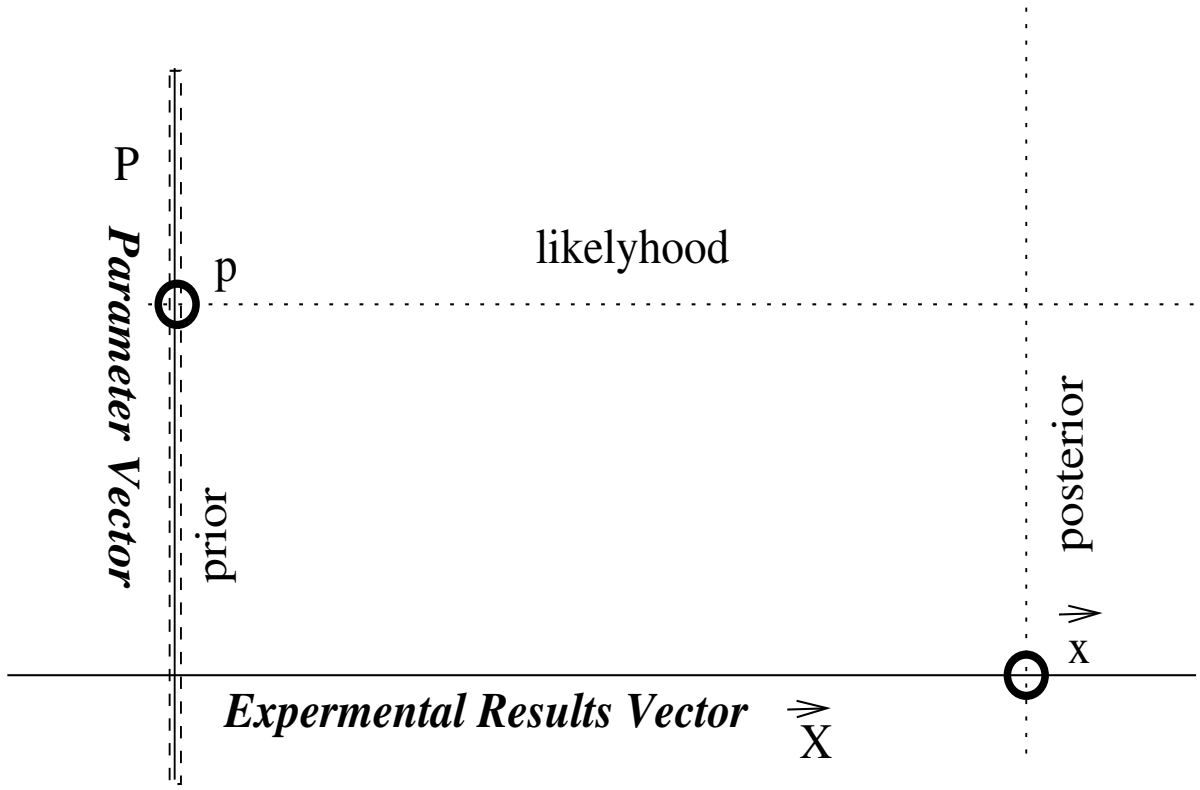
Application: Apply when $l = k = 0$ and explain the relation to our dice rolling experiment.

Example 2 (b): "Clinical Trial" second view

$\vec{\Theta} = P$ continuous with $f_P(p) = \frac{l!k!}{(l+k+1)!} p^k (1-p)^l \chi_{[0,1]}(p)$.

$\vec{X} = (X_1, \ldots, X_N)$ independent Bernoulli trials where

$f_{(\vec{X}|P=p)}(m_1, \ldots, m_N) = p^{(\sum_{i=1}^N m_i)} (1-p)^{(N - \sum_{i=1}^N m_i)}$.



Example 2(b): "Clinical Trial" second view

Hence

$$f_{(P|\vec{X}=(m_1,\ldots,m_N))}(p) \propto p^{(k+\sum_{i=1}^{N} m_i)}(1-p)^{(l+N-\sum_{i=1}^{N} m_i)}\chi_{[0,1]}(p)$$

Explain the relationship between this answer and the answer using the our first view of the Clinical Trial.

Three Key Parameter RVs

One can re-scale the following to make the key the random variables on the $[a, b]$, $(-\infty, b]$, $[a, \infty)$, and $(-\infty, \infty)$ for modeling parameters.

$$f_{N[\mu,\sigma]}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

$$f_{\Gamma[\alpha,\beta]}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \chi_{[0,\infty)}(x)$$

$$f_{\beta[\alpha,\beta]}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \chi_{[0,1]}(x)$$

Some Key Facts

$$\psi_{N[\mu,\sigma]}(t) = E\left(e^{tN[\mu,\sigma]}\right) = e^{\frac{\sigma^2 t^2}{2} + \mu t}$$

$$\sum_{i=1}^{N} N[\mu_i, \sigma_i] = N\left[\sum_{i=1}^{N} \mu_i, \sqrt{\sum_{i=1}^{N} \sigma_i}\right]$$

$$\psi_{\Gamma[\alpha,\beta]}(t) = E\left(e^{t\Gamma[\alpha,\beta]}\right) = \left(\frac{\beta}{\beta - t}\right)^{\alpha}$$

$$\sum_{i=1}^{N} \Gamma[\alpha_i, \beta] = \Gamma\left[\sum_{i=1}^{N} \alpha_i, \beta\right]$$

Three Templates

Let $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\mu_1 = \frac{\sigma^2\mu + n\nu^2\bar{x}}{\sigma^2 + n\nu^2}$, $\nu_1 = \frac{\sigma^2\nu^2}{\sigma^2 + n\nu^2}$.

	Clinical	*Arrival*	*Mean*
prior	$P = \beta[\alpha, \beta]$	$\Lambda = \Gamma[\alpha, \beta]$	$M = N[\mu, \nu]$
likelihood	$\vec{X} = \vec{B}er[p]_n$	$\vec{X} = \vec{P}ois[\lambda]_n$	$\vec{X} = \vec{N}[m, \sigma]_n$
posterior	$\beta[\alpha + n\bar{x}, \beta + n - n\bar{x}]$	$\Gamma[\alpha + n\bar{x}, \beta + n]$	$M = N[\mu_1, \nu_1]$
$E(Prior)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha}{\beta}$	μ
$V(Prior)$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$\frac{\alpha}{\beta^2}$	ν^2
$Loss = (\theta - a)^2$	$\frac{\alpha+n\bar{x}}{\alpha+\beta+n}$	$\frac{\alpha+n\bar{x}}{\beta+n}$	μ_1
$Loss = \lvert\theta - a\rvert$	num	num	μ_1
$MaxLikelihood$	$\bar{x}$	$\bar{x}$	$\bar{x}$

Estimating Parameters

	$Clinical$	$Arrival$	$Mean$
$Natural$	$E(\log(P)), E(\log(1-P))$	$E(\log(\Lambda)), E(\Lambda)$	$E(M), V(M)$
$E(Natural)$	$\Psi(\alpha)-\Psi(\alpha+\beta), \Psi(\beta)-\Psi(\alpha+\beta)$	$\Psi(\alpha)-\log(\beta), \frac{\alpha}{\beta}$	μ, ν
$NoInfoPrior$	$\frac{1}{p(1-p)}\chi_{[0,1]}(p)$	$\frac{1}{\gamma}\chi_{[0,\infty)}(\gamma)$	$\chi_{(-\infty,\infty)}(m)$

Facts About our Estimators

1. If two different sampling plans produce proportional (with respect to θ) likely functions, then the MLEs agree.

(See Exercises 6 and 9 page 334).

2. If $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$. (See the Next Example)

Explore This Example

Example Management at the local Burger Buddy has changed. We intend to examine the service time in minutes required to service a customer under this new management. Suppose that the time in minutes required to service a customer has can be described by an exponential distribution $f_T(t) = \theta e^{-\theta t}$ for which the value of the parameter θ is unknown. (Is this reasonable?)

1. Suppose the average time required to serve a random sample of 30 customers is observed to be 1.6 minutes. How could you estimate θ?

2 (a). Suppose that under the previous management an enormous amount of data was collected and the average service time was found to be 1.6 minutes. How big should your sample size N be so that if the new management really has an expected service time ≤ 1.3 then you are at least 95 percent certain that you will not accidently report a larger expected service time than the previous management's known service time (using your method form 1)? How big should N be so that if the new management really has a service time ≥ 1.9 then we are at least 95 percent certain that we will not accidently report a smaller average service time than the previous management's service time?

3. Suppose you know the average nationwide service time at Burger Buddy is 1.3 minutes. This parameter is very management dependent and usually (about 90 percent of the time) measured to fall within the range of $[1, 2]$ minutes. How might you choose a prior? Does this effect your parameter estimate in 1? What choices did you make? Try out other choices.

Solution to part 1

1. Suppose the average time required to serve a random sample of 30 customers is observed to be 1.6 minutes. How could you estimate θ?

Solution: Let $A_N[\theta] = \frac{\sum_{i=1}^{N} T_i}{N}$ were the T_i are i.i.d. $Exp[\theta]$ random variable representing the customer service times. By the WLLN, this average is an estimate of the expect service time $s = E(T_i)$, and

$$s = E(T_i) = \int_0^\infty \theta e^{-\theta t} dt = \frac{1}{\theta}.$$

So if we could find a Maximum likelihood estimate $\hat{s}$ for s, then by fact 2 on the previous slide $\hat{\theta} = \frac{1}{\hat{s}}$. The likelihood function using s as our parameter is

$$f_{(\vec{T}|S=s)}(\vec{t}) = \Pi_{i=1}^{N} \left(\frac{1}{s} e^{-\frac{t_i}{s}} \right) = \frac{1}{s^N} e^{-\frac{1}{s}\sum_{i=1}^{N} t_i}.$$

Using one variable calculus, we find the maximum is is at $\hat{s} = \frac{\sum_{i=1}^{N} t_i}{N} = 1.6$. So $\hat{\theta} = \frac{1}{\hat{s}} = \frac{1}{1.6} = 0.625$.

Solution to 2(a), Approximate via CLT

2 (a). Suppose that under the previous management an enormous amount of data was collected and the average service time was found to be 1.6 minutes. How big should your sample size N be so that if the new management really has an expected service time ≤ 1.3 then you are at least 95 percent certain that you will not accidently report a larger expected service time than the previous management's known service time (using your method form 1)?

Solution: Recalling that $\theta = 1/E(T_i)$ we are can interpret this request as finding the smallest such N so that

$$\max_{\{\theta \geq 1/1.3\}} Pr(A_N[\theta] \geq 1.6) \leq 0.05$$

The CLT gives us a method to approximate this probability. Since $E(T_i) = 1/\theta$ and $Sd(T_i) = 1/\theta$, for each $\theta \geq 1/1.3$ we are looking for the smallest N such that $1 - F_{N[0,1]}(\sqrt{N}(\theta 1.6 - 1)) \leq 0.05$ and then taking the largest of these Ns. Notice $1 - F_{N[0,1]}(\sqrt{N}(\theta 1.6 - 1)))$ decreases as θ increases, so we may assume $\theta = 1/1.3$ and let $N = ceil\left(\frac{1}{1.6/(1.3)-1} F^{-1}_{N[0,1]}(.95))^2\right) = 51$.

Solution to 2(a), Exact

2 (a). Suppose that under the previous management an enormous amount of data was collected and the average service time was found to be 1.6 minutes. How big should your sample size N be so that if the new management really has an expected service time ≤ 1.3 then you are at least 95 percent certain that you will not accidently report a larger expected service time than the previous management's known service time (using your method form 1)?

Solution: Recall form our key facts $A_N = \frac{1}{N}\Gamma[N, \theta]$. We need the smallest N that guarantees that

$$\max_{\{\theta \geq 1/1.3\}} Pr(\Gamma[N, \theta] \geq 1.6N) \leq 0.05$$

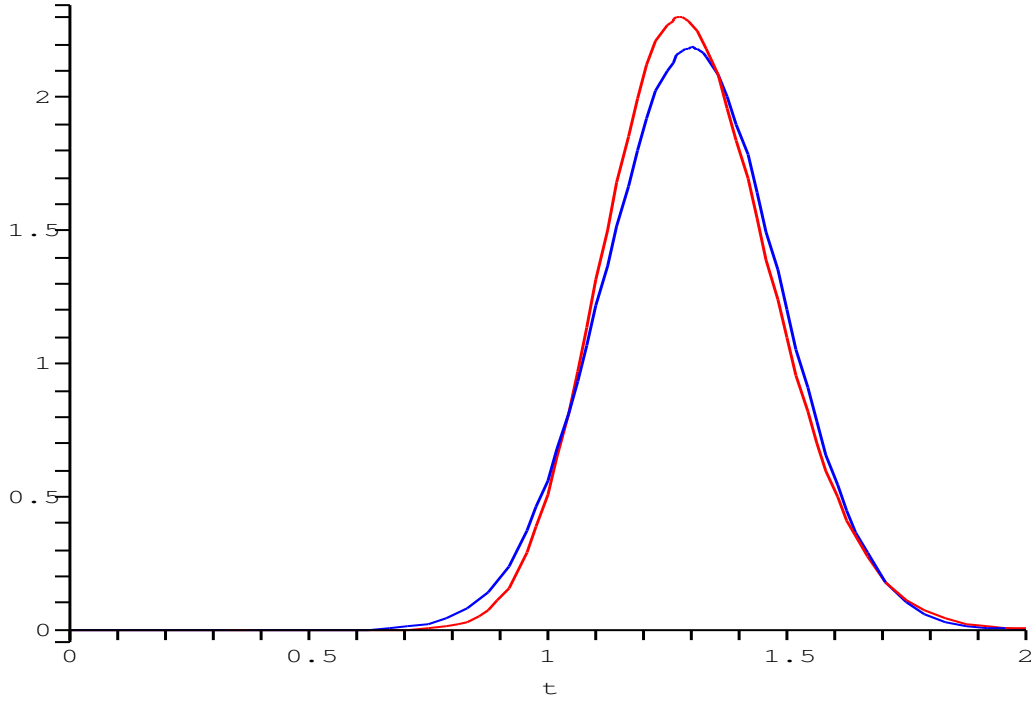
Once again, we'd like to know that this probability decreases as θ increases. To see this we notice

$$\frac{d}{d\theta} Pr(\Gamma[N, \theta] \geq 1.6N) = E((N/\theta - x)\chi_{[1.6N,\infty]}) < E((1.3N - N1.6)\chi_{[1.6N,\infty]}) < 0$$

Now we use maple to find the smallest N such that $Pr(\Gamma[N, \frac{1}{1.3}] \geq 1.6N) \leq 0.05$ to be 56.

A Right Tail.

$56 > 51$ suggest that that $A_N[\theta]$ has right tail. The red curve is actual curve and the blue the curve from the CLT giving us the smallest N such that $Pr(A_N[1/1.3] \geq 1.6) = 0.05$.



Solution 2(b)

2 (b). How big should N be so that if the new management really has a service time ≥ 1.9 then we are at least 95 percent certain that we will not accidently report a smaller average service time than the previous management's service time?

Answer: HW: 109 using the CLT and 101 using the exact method.

Solution 3

3. Suppose you know the average nationwide service time at Burger Buddy is 1.3 minutes. This parameter is very management dependent and usually (about 90 percent of the time) measured to fall within the range of $[1, 2]$ minutes. How might you choose a prior? Does this effect your parameter estimate in 1? What choices did you make? Try out other choices.

Solution: We use the parameter θ and as in part 1 its MLE is $\hat{\theta} = 1/(1.3)$ and we may assume $Pr(\hat{\theta} \in [1/2, 1]) = .9$. Using this information, a natural choice of prior would be $\Gamma[\alpha, \beta](\theta)$ chosen so that $E(\Gamma[\alpha, \beta]) = \frac{1}{1.3}$ and $Pr(\Gamma[\alpha, \beta] \in [1/2, 1]) = .9$. Using maple we find $\alpha = 25.3$ and $\beta = 32.9$. The likelihood function is

$$f_{(\vec{T}|B=b)}(\vec{t}) = \Pi_{i=1}^{N} \left(\theta e^{-\theta t_i}\right) = \theta^N e^{-\theta \sum_{i=1}^{N} t_i},$$

so the posterior is $\Gamma[\alpha + N, \beta + \sum_{i=1}^{N} t_i]$. In particular, using the loss function $L(\theta, a) = (\theta - a)^2$ we would estimate θ to be $\theta^* = \frac{25.3+30}{32.9+30(1.6)} = 0.68$. Hence the expect service time using our Bayesian estimate is 1.46 minutes.

The Prior and Posterior

The red is the prior and the blue the posterior given that we saw an average of 1.6 in 30 trials.