Exam STK 4021/9021 December 2015:
Notes, by Nils Lid Hjort

The Project, Exercise 1

(a) The log-density of the binomial is $y \log p + (m-y)\log(1-p)$, with derivative

$$u(y,p) = \frac{y}{p} - \frac{m-y}{1-p} = \frac{y-mp}{p(1-p)},$$

which has variance $m/\{p(1-p)\}$. The Jeffreys prior is hence proportional to $p^{-1/2}(1-p)^{-1/2}$, which is the Beta$(\frac{1}{2},\frac{1}{2})$.

(b) With Jeffreys priors for p_0 and p_1, we have

$$p_0 \,|\, \text{data} \sim \text{Beta}(\tfrac{1}{2}+y_0, \tfrac{1}{2}+m_0-y_0) \quad \text{and} \quad p_1 \,|\, \text{data} \sim \text{Beta}(\tfrac{1}{2}+y_1, \tfrac{1}{2}+m_1-y_1),$$

by classic results; here $(y_0,m_0) = (22,79)$ and $(y_1,m_1) = (19,78)$. Also, the two parameters are independent. I simulate 10^5 realisations of (p_0,p_1) given data through

$$\texttt{p0sim} = \texttt{rbeta}(\texttt{sim}, 0.5+\texttt{y0}, 0.5+\texttt{m0}-\texttt{y0})$$
$$\texttt{p1sim} = \texttt{rbeta}(\texttt{sim}, 0.5+\texttt{y1}, 0.5+\texttt{m1}-\texttt{y1})$$

and read off histograms or density estimates for $\gamma = \log(p_1/p_0)$ and for $\rho = \exp(\gamma) = p_1/p_0$. With 10^6 simulations, the 0.025, 0.50, 0.975 quantile points are $0.512, 0.877, 1.480$. Hence $\rho = 1$, corresponding to $p_0 = p_1$, is smack in the middle of the soup (the 95% credibility interval is $[0.512, 1.480]$). It appears artificial to claim that there is any noticeable difference.

(c) Since $\log p$ has derivative $1/p$, we have

$$\sqrt{m}(\log \widehat{p} - \log p) \to_d \text{N}(0,\tau^2)$$

with $\tau^2 = (1/p)^2 p(1-p) = 1/p - 1$. This translates to

$$\text{Var} \log \widehat{p} \approx_d \text{N}(0, (1/m)(1/p-1)).$$

The variance here, say κ_m^2, is naturally estimated using $\widehat{\kappa}_m^2 = 1/y - 1/m$. One can easily show that $\widehat{\kappa}_m^2/\widehat{\kappa}_m^2 \to_{\text{pr}} 1$. (Note, incidentally, that $\log \widehat{p}$ is not defined if $y=0$, but this happens with exceedingly low probability.)

(d) From the above we have

$$\log(\widehat{p}_1/\widehat{p}_0) \approx_d \text{N}(0,\sigma^2),$$

say, with

$$\sigma^2 = (1/m_1)(1/p_1-1) + (1/m_0)(1/p_0-1),$$

which we estimate using

$$\widehat{\sigma}^2 = 1/y_1 - 1/m_1 + 1/y_0 - 1/m_0 = 0.269^2.$$

(e) The classic frequentist approximate 95% interval for γ becomes $\widehat{\gamma} \pm 1.96\,\widehat{\sigma}$, which is $[-0.6620, 0.3943]$. Transforming to $\rho = \exp(\gamma)$ scale, the interval is $[0.516, 1.483]$. It is quite close to the Bayesian credibility interval based on Jeffreys, above, as we expect from the general normal approximation theory for posterior distributions.

(f) We learn from putting up prior times likelihood that it takes the form $\exp(-\frac{1}{2}Q)$, with a quadratic function Q in $(\gamma, \widehat{\gamma})$, so the joint distribution is binormal. It has the mean and variance matrix structure as given in the exercise. Here we use e.g. that

$$\operatorname{Var}\widehat{\gamma} = \operatorname{E}\operatorname{Var}(\widehat{\gamma}\,|\,\gamma) + \operatorname{Var}\operatorname{E}(\widehat{\gamma}\,|\,\gamma) = \operatorname{E}\sigma^2 + \operatorname{Var}\gamma = \sigma_0^2 + \sigma^2.$$

The conditional distribution

$$\gamma\,|\,\widehat{\gamma} \sim \mathrm{N}(\widehat{\gamma}_B, \sigma_B^2),$$

with formulae as in the exercise, follows from the appropriate Nils Exercises.

(g) Suppose $\gamma \sim \mathrm{N}(0, \sigma_0^2)$ with $\sigma_0 = 0.35$ (which is Laptook et al.'s 'neutral prior'). The central 95% prior credibility interval is $\gamma_0 \pm 1.96\,\sigma_0 = \pm 0.686$, which translates to the interval $[0.504, 1.986]$ on the ρ scale. The prior density is log-normal,

$$f_0(\rho) = \phi\Big(\frac{\log\rho - \gamma_0}{\sigma_0}\Big)\frac{1}{\sigma_0}\frac{1}{\rho} \quad \text{for } \rho > 0,$$

with $\phi(\cdot)$ the standard normal density.

(h) Combining this with $\widehat{\gamma}\,|\,\gamma \approx \mathrm{N}(\gamma, \widehat{\sigma}^2)$, with $\widehat{\sigma}$ well estimated as 0.269, leads by the above to

$$\gamma\,|\,\widehat{\gamma} \approx_d \mathrm{N}(-0.084, 0.214^2).$$

So the posterior for γ is approximately normal, and by easy transformation the posterior for ρ is approximately

$$f(\rho) = \phi\Big(\frac{\log\rho - \widehat{\gamma}_B}{\widehat{\sigma}_B}\Big)\frac{1}{\widehat{\sigma}_B}\frac{1}{\rho} \quad \text{for } \rho > 0.$$

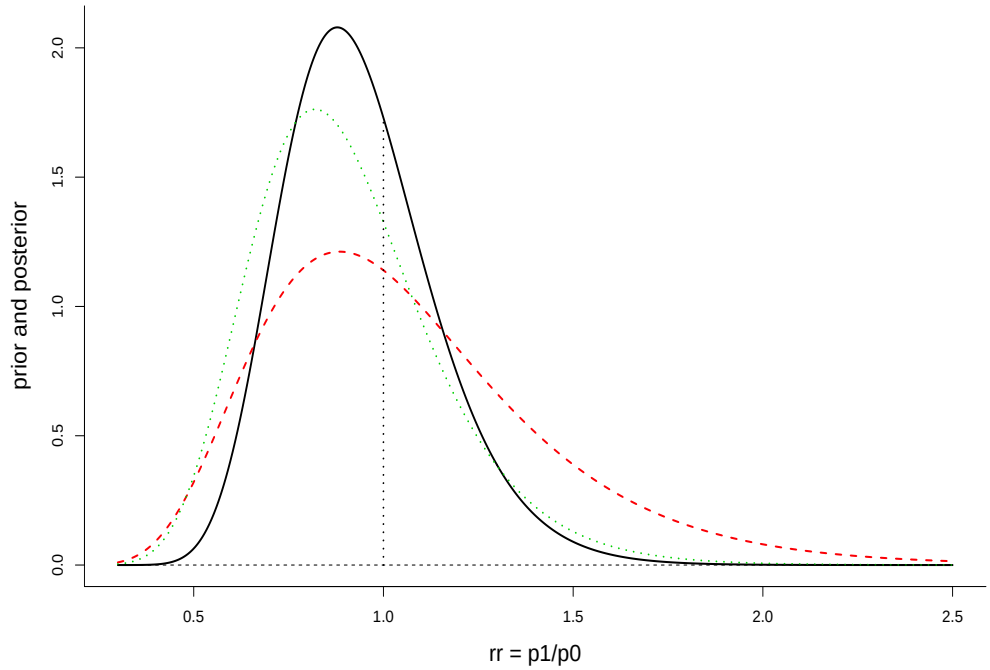
The posterior density for ρ with the Jeffreys priors, used above, can be well estimated via a histogram, or using a kernel density estimator, with say 10^6 simulations. A precise curve can also be computed numerically, however, bypassing approximations and simulations. Using g_0 and g_1 for the posterior Beta densities of p_0 and p_1, with cumulatives G_0 and G_1, the posterior cumulative for ρ is

$$H(\rho) = P\{p_1 \le \rho p_0\,|\,\text{data}\} = \int G_1(\rho p_0) g_0(p_0)\,\mathrm{d}p_0$$

with derivative, i.e. the posterior density, equal to

$$h(\rho) = \int g_1(\rho p_0) p_0 f_0(p_0)\,\mathrm{d}p_0.$$

This can be evaluated via numerical integration, using `dbeta` and indeed `integrate`. The figure displays the neutral prior for ρ (red, slanted); the posterior based on neutral prior and using the normal approximation above (black, full), which presumably what Laptook et al. have done; and the exact posterior density based on the Jeffreys priors (green, dotted). Again, the message is that $\rho = 1$ is smack in the middle, with no clear evidence that we can claim $\rho < 1$.



(i) I'm computing

$$Q(y_1) = P\{p_1 < p_0 \,|\, y_0 = 22, y_1\}$$

using the Jeffreys priors, and for each value $y_1 = 7, 8, 9, \ldots, 18, 19$. This can be done via simulations, simulating for each value of y_1 a high number of (p_0, p_1) from the posterior distribution and then reading off the proportion of $p_1 < p_0$ cases. It may also be done via exact numerical integration, however:

$$Q(y_1) = \int G_1(p_0 \,|\, y_1) g_0(p_0 \,|\, y_0)\, \mathrm{d}p_0,$$

wth $g_0(p_0 \,|\, y_0)$ being the Beta$(0.5 + y_0, 0.5 + m_0 - y_0)$ density and $G_1(x \,|\, y_1)$ being the Beta$(0.5 + y_1, 0.5 + m_1 - y_1)$ cumulative. The result is a nice little plot $(y_1, Q(y_1))$, with $Q(19) = 0.690$ for the Laptook data (far too low, to really claim that $p_1 < p_0$), and $Q(13) = 0.954$, and even higher, of course, for y_1 even smaller than 13. So from the traditional point of view, one would claim signficance only with $y_1 \leq 13$; $y_1 = 19$ is not at all statistically impressive.

(j) Laptook et al. work with their 'neutral prior' where $\gamma \sim \mathrm{N}(0, 0.35^2)$, which we have seen corresponds to $\rho \in [0.504, 1.986]$ as the natural 95% prior credibility interval

(with median 1, rather than mean 1; the main figure 2.1 in Laptook et al. (2017) appears to be wrong. Finding Beta priors for (p_0, p_1) to match mean 0 and variance 0.35^2 for γ can be carried out in several ways. The easiest is to let both p_0 and p_1 come from the same Beta(a_0, b_0), which ensures $\log(p_1/p_0)$ having mean zero, and fixing (a_0, b_0) to have variance

$$\text{Var } \log p = \int_0^1 (\log p)^2 g(p \,|\, a_0, b_0)\, \mathrm{d}p - \Big\{ \int_0^1 \log p \, g(p, a_0, b_0)\, \mathrm{d}p \Big\}^2$$

half of 0.35^2. There are again many such Beta densities, but we may e.g. have it centred in 0.25, which means we should find k such that the the logarithm of Beta$(k\,0.25, k\,0.75)$ has variance precisely equal to $\frac{1}{2}\,0.35^2 = 0.06125$. Computing this and fine-tuning for k leads to $k = 51.421$. In other words, if $p \sim \text{Beta}(a_0, b_0)$ with $(a_0, b_0) = (12.855, 38.566)$, then its logarithm has the right variance. – You may draw the corresponding prior density, in this setup common to both p_0 and p_1, to see how 'neutral' the Laptook et al. neutral prior is.

(k) I do not embellish things here; have rather a look at my FocuStat blog post. But, briefly:

[i] There's really no serious statistical signpost here saying p_1 is smaller than p_0; both the frequentist and Bayesian methods lead to confidence and credibility intervals havint the value $\rho = p_1/p_0 = 1$ smack in the middle.

[ii] This is a bigger discussion, not easy to sort out well. There is a role for Bayes and priors in clinical trials, but care needs to be exercised. And in case where the frequentist analysis is clear, as here, there's a risk that Bayes and priors might lead to confusion. Again, see my blog post, *Cooling of Newborns and the Difference Between 0.244 and 0.278*.

The Project, Exercise 2

(a) Taking the logarithm of the likelihood function

$$\prod_{j=1}^{10} \binom{m}{y_j} p_j(a,b)^{y_j} \{1 - p_j(a,b)\}^{m-y_j}$$

leads to the log-likelihood function

$$\ell(a,b) = \sum_{j=1}^{10} \Big[y_j \log p_j(a,b) + (m - y_j) \log\{1 - p_j(a,b)\} + \log \binom{m}{y_j} \Big].$$

I've programmed `logL` in R, as a function of `ab`, complete with `a=ab[1], b=ab[2]`, etc. Via `nlm` for `minuslogL` I find ML estimates $(-1.479, 0.949)$. Standard errors, via the inverse observed Fisher information matrix, are $(0.690, 0.546)$.
With the right plotting tools I then construct the figure given in the project.

(b) I set up a MCMC scheme in the same fashion I've used for some exercises in the course, giving me a `sim` $\times 2$ matrix for a chain of (a, b) outcomes. I use an easy normal proposal, of the symmetric type

$$(a', b') = (a + \varepsilon_1, b + \varepsilon_2),$$

with $\varepsilon_1 \sim \mathrm{N}(0, \sigma_1^2)$ and $\varepsilon_2 \sim \mathrm{N}(0, \sigma_2^2)$. In principle there's a broad range of σ_1 and σ_2 which will work. I've used $\sigma_1 = \frac{1}{2}\, 0.690$ and $\sigma_2 = \frac{1}{2}\, 0.546$, half of the standard errors for the ML estimators (but, again, lots of other schemes will work).

– The posterior distribution for (a, b) obtained in this way is fairly binormal, with no drama. For a (first row) and b (second row) I record ML and standard error, then mean and standard deviation for the posterior distribution:

$$\begin{aligned} -1.479, 0.690 \qquad & -1.567, 0.713 \\ 0.949, 0.546 \qquad & 1.008, 0.554 \end{aligned}$$

There is clear and strong agreement, so we are in the Good Terrain of Bernshteĭn–von Mises, as in the book's Chapter 4.

(c) I then used simulations to read off the 0.05 quantile $\mathrm{low}(x)$ and 0.95 quantile $\mathrm{up}(x)$, from the posterior distribution of $H(a + bx)$. In other words, first I got hold of 10^5 values of (a_j, b_j), via the MCMC. The I read off, for each given x, the consequent 10^5 numbers $H(a_j + b_j x)$. From these I read off the 0.05 and 0.95 quantiles, via `quantile`.

(d) I can also generate 10^5 values of $p(x_{\mathrm{new}}) = H(a_j + b_j x_{\mathrm{new}})$, look at a histogram, etc. The histogram is skewed to the left. Quantiles for 0.05, 0.50, 0.95 are 0.415, 0.719, 0.907. This gives good assessment for $p(x)$ at this new position, provided the model assumption still holds, which might not be the case; cf. the Aukrust brothers.

(e) I generate 10^5 values of (a_j, b_j) via the MCMC; then, for each of these pairs, I generate y_j from the binomial (m, p_j), where $p_j = H(a_j + b_j x_{\mathrm{new}})$. This gives the predictive distribution for y_{new}. From 10^5 simulations, I read off probabilities

$$0.0139, 0.0569, 0.1369, 0.2403, 0.3100, 0.2420$$

for values 0, 1, 2, 3, 4, 5.

(f) Now the restriction is $b \geq 0$ for the (a, b) pair. I've set up a Metropolis MCMC where I make sure that if an (a, b) is proposed with negative b, then that proposal has zero chance of being accepted. This leads to a new MCMC, again of size $\mathrm{sim} \times 2$, and with every (a_j, b_j) respecting $b_j \geq 0$. The posterior distribution is not binormal anymore, with a sharp boundary $b = 0$. I can read off $x_{0,j} = -a_j/b_j$, the LD50 parameter. I can display a histogram and compute quantiles. The 0.05, 0.50, 0.95 quantiles of the posterior distribution for x_0 are 1.051, 1.545, 3.114.

This is a parameter where the usual normal theory does not work well, but where the Bayesian scheme of things work out nicely (as long as the MCMC is set up correctly).

The four-hour exam

Exercise 1

(a) First, $y_i \,|\, \theta$ is a binomial $(1, \theta)$, with mean θ and variance $\theta(1-\theta)$. Secondly, $z = \sum_{i=1}^{n} y_i$ given θ is binomial (n, θ), hence with mean and variance $n\theta$ and $n\theta(1-\theta)$.

(b) Now $\theta \sim \text{Beta}(2, 2)$.

(i) The mean is $\frac{1}{2}$, the variance is $\frac{1}{2}\frac{1}{2}/5 = 1/20$.

(ii) $\mathrm{E}\, y_i = \mathrm{E}\,\mathrm{E}\,(y_i \,|\, \theta) = \mathrm{E}\,\theta = \frac{1}{2}$, and

$$\begin{aligned} \mathrm{Var}\, y_i &= \mathrm{E}\,\mathrm{Var}\,(y_i \,|\, \theta) + \mathrm{Var}\,\mathrm{E}\,(y_i \,|\, \theta) \\ &= \mathrm{E}\,\theta(1-\theta) + \mathrm{Var}\,\theta \\ &= \tfrac{1}{2} - \mathrm{E}\,\theta^2 + \mathrm{E}\,\theta^2 - \tfrac{1}{4} = \tfrac{1}{4}. \end{aligned}$$

Also, y_i is a 0-1 variable, with $P(y_i = 1) = \frac{1}{2}$; hence its variance must be $\frac{1}{2}\frac{1}{2} = \frac{1}{4}$.

(iii) We have $\mathrm{E}\,(y_i y_j \,|\, \theta) = \theta^2$ and hence $\mathrm{cov}(y_i, y_j) = \mathrm{E}\,\theta^2 - (\frac{1}{2})^2 = \mathrm{Var}\,\theta$, i.e. $1/20$. Hence the correlation is

$$\rho = \frac{1/20}{1/4} = \frac{4}{20} = 1/5 = 0.20.$$

(iv) We have

$$\mathrm{E}\, z = \mathrm{E}\, n\theta = \tfrac{1}{2}n$$

and

$$\mathrm{Var}\, z = \mathrm{E}\, n\theta(1-\theta) + \mathrm{Var}\, n\theta = n(\tfrac{1}{2} - \tfrac{1}{4} - 1/20) + n^2/20 = 0.20\, n + 0.05\, n^2.$$

(c) The distribution of z is

$$\begin{aligned} f(z) &= \int_0^1 \binom{n}{z} \theta^z (1-\theta)^{n-z} 6\theta(1-\theta)\, \mathrm{d}\theta \\ &= \frac{n!}{z!\,(n-z)!} \frac{(z+1)!\,(n-z+1)!}{(n+2)!} \\ &= 6\frac{(z+1)(n-z+1)}{n(n+1)} \end{aligned}$$

for $z = 0, 1, \ldots, n$.

(d) The posterior is proportional to

$$\theta^z (1-\theta)^{n-z} \theta(1-\theta) = \theta^{z+1}(1-\theta)^{n-z+1},$$

which is a $\text{Beta}(z+2, n-z+2)$. The conditional mean is

$$\widehat{\theta}_B = \mathrm{E}\,(\theta \,|\, z) = \frac{z+2}{n+4}.$$

(e) The standard estimator is $\widetilde{\theta} = z/n$, which is unbiased. Hence the risk, the mean of $(\widetilde{\theta} - \theta)^2$, is $\theta(1-\theta)/n$. It starts and ends at zero, and is biggest for $\theta = \frac{1}{2}$.

(f) The Bayes estimator has risk function

$$\begin{aligned} r(\theta) &= \operatorname{Var} \widehat{\theta}_B + (\operatorname{E} \widehat{\theta}_B - \theta)^2 \\ &= \frac{n\theta(1-\theta)}{(n+4)^2} + \Big(\frac{n\theta+2}{n+4} - \frac{n\theta+4\theta}{n+4}\Big)^2 \\ &= \frac{n}{(n+4)^2}\theta(1-\theta) + \frac{(4\theta-2)^2}{(n+4)^2}. \end{aligned}$$

It is better than the usual ML estimator for those θ where

$$\frac{16(\theta-\frac{1}{2})^2}{(n+4)^2} \le \Big\{\frac{1}{n} - \frac{n}{(n+4)^2}\Big\}\theta(1-\theta),$$

or

$$16(\theta-\tfrac{1}{2})^2 \le \frac{(n+4)^2-n^2}{n}\theta(1-\theta) = \frac{16+8n}{n}\theta(1-\theta),$$

or $(\theta-\frac{1}{2})^2 \le (\frac{1}{2}+1/n)\theta(1-\theta)$. This means a certain interval around $\frac{1}{2}$, where Bayes is better. For large n, the inequality is close to $(\theta-\frac{1}{2})^2 \le \frac{1}{2}\theta(1-\theta)$, and this holds for θ inside $\frac{1}{2} \pm \sqrt{3}/6$, which is $[0.211, 0.789]$, i.e. a pretty wide interval.

Exercise 2

(a) The likelihood function becomes

$$\begin{aligned} L &= (p(1-q))^{212}\,((1-p)q)^{103}\,(pq)^{39}\,((1-p)(1-q))^{148} \\ &= p^{212+39}(1-p)^{103+148}q^{103+39}(1-q)^{212+148} \\ &= p^{251}(1-p)^{251}q^{142}(1-q)^{360}. \end{aligned}$$

(b) With independent uniforms for p and q, the posteriors are also independent, with

$$p \,|\, \text{data} \sim \text{Beta}(252, 252), \quad q \,|\, \text{data} \sim \text{Beta}(143, 361).$$

(c) The expected number of AB cases, if this theory is correct, is

$$e_{AB} = \operatorname{E}(npq \,|\, \text{data}) = n\frac{252}{502}\frac{143}{502} = 71.5.$$

But this is far off from the observed 39. So the theory looks very suspicious, indeed. As Landsteiner and others found out, about a hundred years ago, the two-loci theory stinks and sucks; the one-locus theory, however, is splendid. – One may put in more detail here, including computing the probability that one should get a number as far off as 39 (or more), as measured through the lens of the posterior distribution for npq. This will be a microscopic probability. The essence is simply to compare the observed

far too small 39 with the mean of 71.5 – and, of course, similar calculations for the other three cells.

Exercise 3

(a) The posterior density $p(\theta)$ is the derivative of the cumulative $P(\theta)$, and one finds $\theta \exp(-\theta)$. This is also a gamma $(2, 1)$. Its mean is 2. The density is zero at zero, climbs to $\exp(-1) = 0.368$ at the value 1, and then decreases slowly to zero.

(b) The Bayes decision is to pick among A, B, C the action that has the smallest expected posterior loss. These three expected posterior losses are

$$\begin{aligned} 2\,\{1 - P(1.1)\} &= 2 \cdot 0.6990 = 1.3981, \\ 3\,\{P(1.1) + 1 - P(3.3)\} &= 3 \cdot 0.4596 = 1.3787, \\ 4\,P(3.3) &= 4 \cdot 9.8414 = 3.3656, \end{aligned}$$

for respectively A, B, C. So we take action B.

(c) The risk function, for a credibility interval $[\widehat{a}, \widehat{b}]$ constructed from the data, becomes

$$r(\theta) = \mathrm{E}_\theta\, L(\theta, [\widehat{a}, \widehat{b}]) = 0.10\, E_\theta\, (\widehat{b} - \widehat{a}) + 1 - P_\theta\{\theta \in [\widehat{a}, \widehat{b}]\}.$$

A good method is one with short expected length and with high probability of containing the right parameter.

(d) Again we ought to minimise posterior expected loss. This means minimising

$$\mathrm{E}\,\{L(\theta, [a, b]) \,|\, \text{data}\} = 0.10\,(b - a) + P(a) + 1 - P(b)$$

over (a, b). Taking derivatives leads to the equations $p(a) = 0.10$, $p(b) = 0.10$. This again means finding a to the left of 1 and b to the right of 1, as solutions to $p(x) = 0.10$. I find $[a, b] = [0.112, 3.576]$.

Exercise 4

(a) The likelihood for the data becomes

$$\prod_{i=1}^{n} \{(3\theta y_i^2) \exp(-\theta y_i^3)\} \propto \theta^n \exp(-n\theta W_n),$$

where $W_n = (1/n) \sum_{i=1}^n y_i^3$. Its logarithm is

$$\ell_n(\theta) = n \log \theta - n\theta W_n,$$

with first derivative $n/\theta - nW_n$ and second derivative $-n/\theta^2$. So the maximum likelihood (ML) estimator is

$$\widehat{\theta} = 1/W_n.$$

(b) The prior times the likelihood is proportional to

$$\theta^{a-1}\exp(-b\theta)\theta^n\exp(-n\theta W_n) = \theta^{a+n-1}\exp\{-(b+nW_n)\theta\},$$

which means the posterior is a Gamma with parameters $(a+n, b+nW_n)$. Its mean, by the way, is the Bayes estimator

$$\widehat{\theta}_B = \frac{a+n}{b+nW_n},$$

which is close to the ML estimator.

(c) First, by frequentist ML theory, and assuming the model is actually correct,

$$\widehat{\theta} \approx_d \mathrm{N}(\theta_0, \theta_0^2/n),$$

with θ_0 signalling the true parameter value. Secondly, by Bayes theory for larger sample sizes,

$$\theta \,|\, \text{data} \approx_d \mathrm{N}(\widehat{\theta}, \widehat{\theta}^2/n).$$

So there's a mirror situation, and the Bayesian and the frequentist will have the same inferences, for large n.