STAT 206
Take-Home Test 1

(1) A) True. Using the classical (Pascal-Fermat) interpretation of probability, each outcome on the wheel is regarded as equipossible because of the symmetry argument. Therefore, the probability is equal to

$$\frac{\text{number of red outcomes}}{\text{\# of possible outcomes}} = \frac{18}{38}$$

B) True. We are given that the wheel is fair and each outcome is equally likely. Therefore, we can define the probability of the wheel landing on a red number as the "relative frequency obtained in a long sequence of tosses [trials] assumed to be performed in an identical manner" (Gelman et al). Since the wheel is fair, the probability of spinning a red number will converge to that of the classical interpretation ($^{18}/_{38}$) over many trials.

CA) False. When we take away the fair assumption of the wheel spins we cannot assume 'symmetry' or 'exchangeability' of outcomes. Therefore, we don't know that each outcome is equipossible and cannot use the Classical interpretation to define the probability of spinning a red number.

CB) False. When we don't know that the outcomes are equipossible, our frequentist estimation of probability cannot be estimated using a limit function. Instead, we have to conduct many identical, independent trials to eliminate uncertainty, and until then it cannot be concluded that P(Red #) = $^{18}/_{38}$

D) False. By disposing of our vector y, we lose the ability to quantify the uncertainty of our estimation of θ. We may still calculate an estimate for probability by dividing S_n (sum of trials) by n (number of trials), but we can no longer find the probability distribution.

E) False. This statement is backwards; instead the frequentist approach to probability ensures logical internal consistency of your uncertainty assessments but does not guarantee good calibration, and the Bayesian approach to probability provides a natural framework in which to see if your frequentist answer is well-calibrated.

F) True. Beta distributions are useful as a prior for "random behavior of percentages and proportions" (Wikipedia: Beta Distributions). Distributions on $(0, 1)$ follow this pattern as we can define 0 as 0% and 1 as 100%.

G) False. Defining our model is not easy because there are many ways to specify these ingredients. Rather than forming a hypothesis from the start and testing it, we are able to test several different hypotheses and update their probabilities depending on the information we include.

H) True. If a proposition B_i has no "credence" (From Wikipedia: Credence (statistics) is a "measure of belief strength, expressed as a percentage." If a piece of evidence has no credence (i.e. it is false) its belief strength would be zero and using Bayes' theorem would result in dividing by zero.

$$P(A|B_i) = \frac{P(B_i|A) \cdot P(A)}{P(B_i) \leftarrow \text{zero.}}$$

I) True. Objective probability is defined under the frequentist framework. However, not everyone who makes the probability assessment will agree on this framework.

J) False. We cannot maximize the utility function $U(a, \theta | B)$ when θ is an unknown because that means our utility function is a random variable. We can maximize the expected utility, by using estimates for θ.

K) True. Laplace's technique for approximating higher dimensional integrals was only rediscovered in the 1950s.

L) False. To correct this statement, epistemology and Ontology must swap places. The former relates to the study of knowledge, whereas the latter relates to the study of being.

STAT 206

Take-Home Test 1: Part 2 (Calculation)

(A)

		Truth: HIV ⊕ ($\theta=1$)	Truth: HIV ⊖ ($\theta=0$)	Total
Blood Test	⊕ ($y_1=1$)	TP: $\alpha\beta$	FP: $(1-\alpha)(1-\gamma)$	$\alpha\beta+(1-\alpha)(1-\gamma)$
	⊖ ($y_1=0$)	FN: $\alpha(1-\beta)$	TN: $(1-\alpha)\gamma$	$\alpha(1-\beta)+(1-\alpha)\gamma$
Total		α	$1-\alpha$	1

i) $\beta = P(y_1=1 \mid \theta=1, B)$ $\quad \alpha = P(\theta=1 \mid B)$

a. $TP = P(y_1=1 \text{ and } \theta=1 \mid B)$

$\therefore \equiv \alpha\beta$ using the multiplication rule.

b. $TN = P(y_1=0, \theta=0 \mid B)$

$1-\alpha = P(\theta=0 \mid B)$ $\quad \gamma = P(y_1=0 \mid \theta=0, B)$

$TN \equiv \gamma(1-\alpha)$

c. $FN = P(y_1=0, \theta=1 \mid B) = \alpha(1-\beta)$

$1-\beta = P(y_1=0 \mid \theta=1, B)$ $\quad \alpha = P(\theta=1 \mid B)$

False negatives are the proportion of remaining HIV⊕ ($\theta=1$) patients since the blood test (y_1) is binary.

d. $FP = P(y_1=1, \theta=0 \mid B) = (1-\alpha)(1-\gamma)$

$1-\alpha = P(\theta=0 \mid B)$ $\quad 1-\gamma = P(y_1=1 \mid \theta=0, B)$

$\therefore$

False positives are the remaining proportion of HIV⊖ ($\theta=0$) patients who test positive ($y_1=1$). Because y_1 is binary, the probability $P(y_1=1 \mid \theta=0, B)$ can be represented as $1-\gamma$ or $1-P(y_1=0 \mid \theta=0, B)$.

ii) Use the table to write down formulas for PPV (precision), and NPV in terms of (α, β, γ).

$\alpha = P(\theta=1 \mid B)$
$\beta = P(y_1=1 \mid \theta=1, B)$
$\gamma = P(y_1=0 \mid \theta=0, B)$

$PPV = P(\theta=1 \mid y_1=1, B)$

with our given probabilities, we can calculate PPV using Bayes' Theorem

$$P(\theta=1 \mid y_1=1, B) = \frac{\beta \cdot \alpha}{P(y_1=1 \mid B)}$$

$\beta \cdot \alpha + (1-\alpha)(1-\gamma)$

∴ using the table

$$\boxed{PPV = \frac{\beta \cdot \alpha}{[\alpha\beta + (1-\alpha)(1-\gamma)]}}$$

$NPV = P(\theta=0 \mid y_1=0, B)$

$$\boxed{NPV = \frac{\gamma(1-\alpha)}{\alpha(1-\beta)+(1-\alpha)\gamma}}$$

$[1-\alpha = P(\theta=0 \mid B)]$
denominator $\equiv P(y_1=0 \mid B)$

FDR = 1 − PPV

$FDR = \frac{FP}{FP+TP}$ is related to PPV in that they share the same denominator. However, PPV uses true positives for the numerator because we are measuring how often a person who tests positive for HIV, truly has the condition.

FOR = 1 − NPV

$FOR = \frac{FN}{FN+TN}$ shares the same denominator as NPV. NPV uses true negatives in the numerator since we are measuring the rate at which a negative blood test corresponds to a patient who truly doesn't have HIV.

iii) $\alpha^* = 0.00286 \quad (\beta, \gamma) = (0.99, 0.95)$

Calculate PPV, NPV, FDR, and FOR.

$$\widehat{PPV} = \frac{\beta \cdot \alpha^*}{\alpha^*\beta + (1-\alpha^*)(1-\gamma)} = \frac{0.99(0.00286)}{0.00286(0.99) + (1-0.00286)(0.05)}$$

$$= \frac{0.00028314}{0.0544} = \boxed{0.0537 \quad \widehat{PPV}}$$

$$\widehat{NPV} = \frac{\gamma(1-\alpha^*)}{\alpha^*(1-\beta) + (1-\alpha^*)\gamma} = \frac{0.95(0.99714)}{0.00286(0.01) + (0.99714)(0.95)}$$

$$\widehat{NPV} = \frac{0.947283}{0.9473116} = 0.99997$$

$$\widehat{FDR} = 1 - \widehat{PPV} = 1 - 0.0537$$

$$\widehat{FDR} = 0.9463$$

$$\widehat{FOR} = 1 - \widehat{NPV} = 1 - 0.99997$$

$$\widehat{FOR} = 0.00003$$

Using our calculated False Omission Rate ($\widehat{FOR}$) we find that there is about a 0.003% chance that someone who has HIV would test negative for it using the E_1 test.

Was the Red Cross successful at keeping HIV out of the blood supply?

↓↓

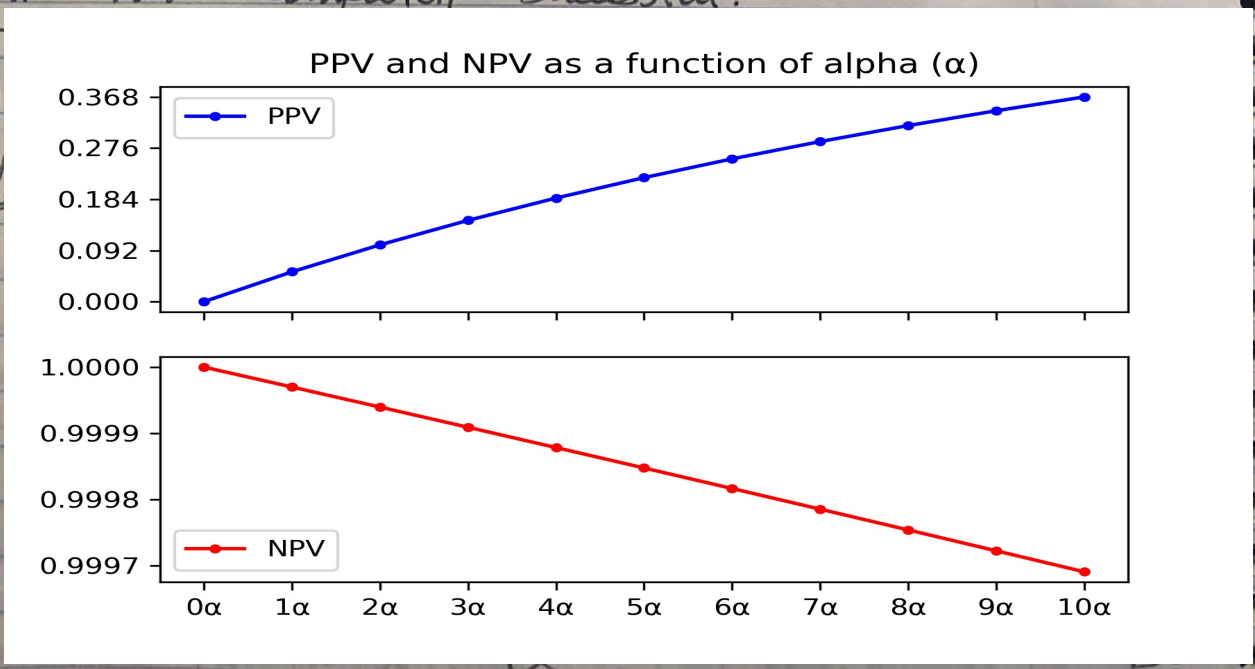
Some factors to consider:

- How likely is someone who has HIV going to attempt to donate blood? Our FOR says that 1 in ~33,000 people with HIV will test negative. If there are ~500,000 people with HIV in the U.S. and say 3% of them try to donate (redcross.org, only an estimate) then we are left with 15,000 people. Quite possible one slips through.
- Wouldn't people who know they have HIV be less likely to try to donate blood?
- Is "highly successful" only if none of the donors have HIV? What are we comparing the E_1 test to?

I would say the E_1 test was highly successful, but not completely successful.

iv)

PPV



PPV is highly sensitive to the prevalance of HIV. If we have 10x the prevalence of λ^*, our Precision (PPV) goes up by about 7x.

On the other hand, our NPV value changes very little as the HIV prevalence goes up (0.9997 at α^* versus 0.9997 at $10\alpha^*$)

<u>However</u>, this is still a very significant change. Rather than 1 in ~33,000 HIV ⊕ patients wrongly testing negative, it would now be 1 in ~3,300; 10 times more likely to occur.

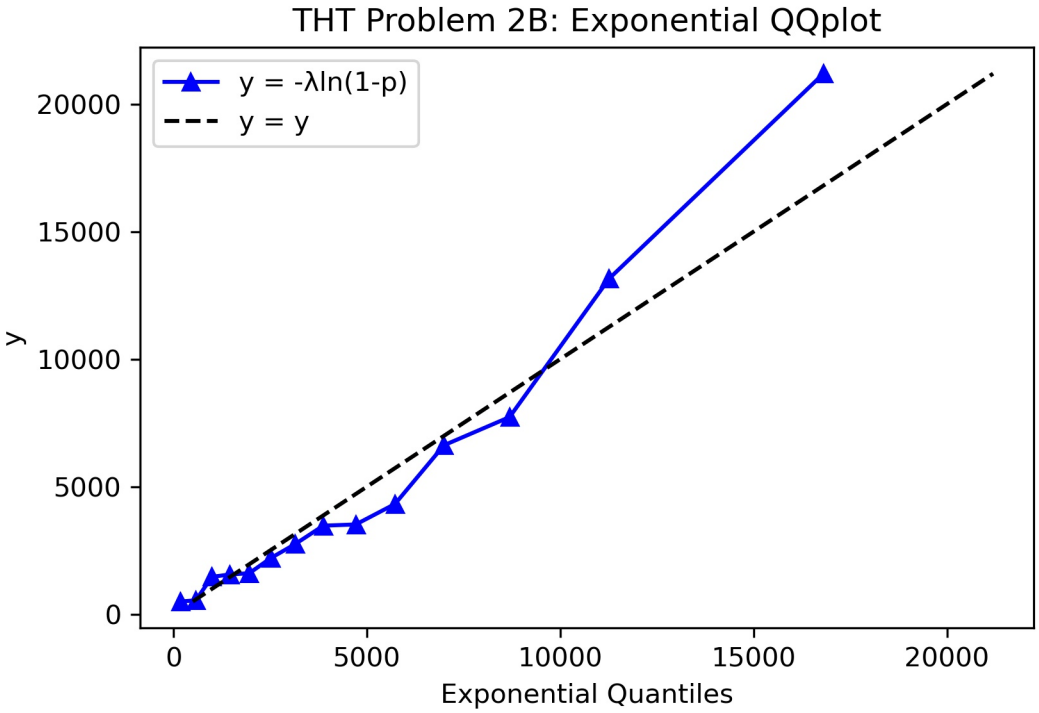
V)

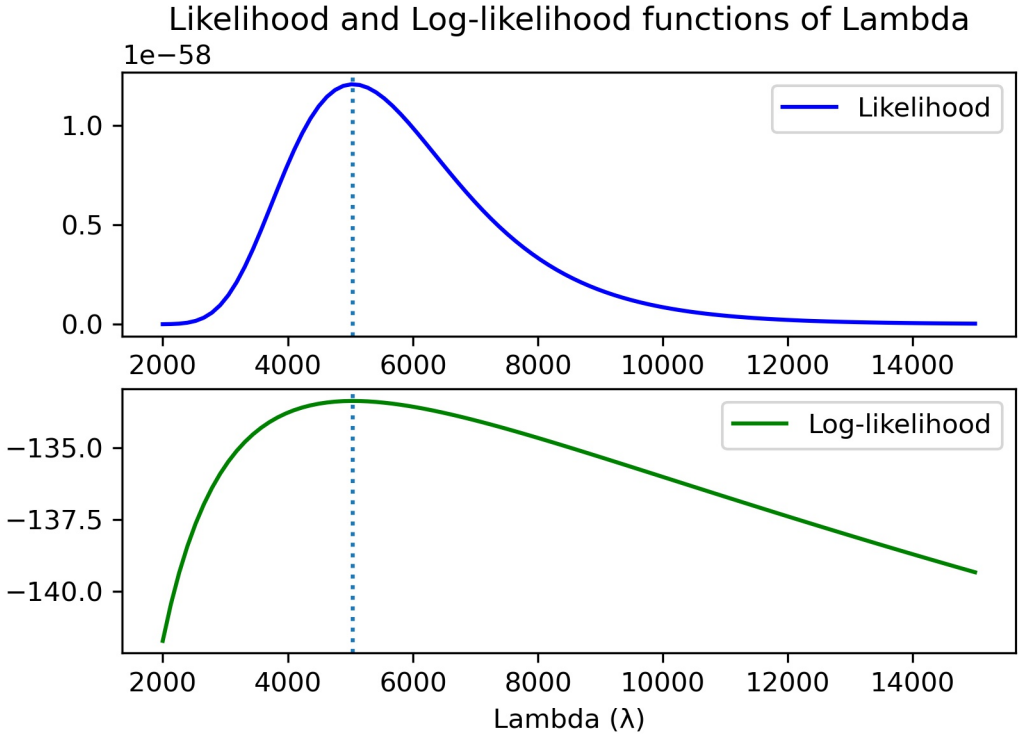
		Truth: HIV ⊕ ($\theta=1$)	HIV ⊖ ($\theta=0$)	Total
Blood Test	⊕ ($Y_1=1$)	495,000	8,725,000	9,220,000
	⊖ ($Y_1=0$)	5,000	165,775,000	165,780,000
	Total	500,000	174,500,000	175,000,000

We can obtain a table of expected cell and margin counts by substituting in our (α, β, γ) values and multiplying each cell by N because none of our cells in table 1 are conditional. This means each of the probabilities are represented as a proportion of the entire population rather than subgroups.

If every citizen were to take the E_1 test, there would be almost 9 million people who would test positive despite not truly having the condition. This would cause a lot of fear or stigma against these individuals. With over 90% of the individuals who test positive being people who don't have the disease, I do not think it would be a good idea to forcibly administer this test.

Data visualizations for 2B





(B) a) $F_Y(y|\lambda E) = \frac{1}{\lambda}\int_{0^+}^{y} e^{-y_i/\lambda}\,dy_i$ $\quad u = -y_i/\lambda$, $du = -\frac{1}{\lambda}dy_i$

$= \int_{0^+}^{\infty} -e^u\,du = \left[-e^{-y_i/\lambda}\right]_{0^+}^{y} = 1 - e^{-y/\lambda}$

$p = 1 - e^{-y/\lambda}$ $\therefore$ to get $F^{-1}(p|\lambda E)$ we isolate y

$e^{-y/\lambda} = 1-p$

$\ln(e^{-y/\lambda}) = \ln(1-p)$

$-y/\lambda = \ln(1-p)$ $\therefore y = F^{-1}(p|\lambda E)$

$-y = \lambda \ln(1-p)$

$= -\lambda \ln(1-p)$

b) $L(\lambda) = \frac{1}{\lambda^n} e\left(\frac{-\sum_{i=1}^{n} y_i}{\lambda}\right)$

$LL(\lambda) = \ln(L(\lambda)) = -n\ln(\lambda) - \frac{1}{\lambda}\sum_{i=1}^{n} y_i$

To get MLE of λ, $\frac{dLL(\lambda)}{d\lambda}(LL(\lambda)) = 0$

$0 = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{n} y_i$

$\frac{n}{\lambda} = \frac{1}{\lambda^2}\sum_{i=1}^{n} y_i$

$\frac{n\lambda^2}{\lambda} = \sum_{i=1}^{n} y_i$

$\hat{\lambda}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$

Our MLE for $\hat{\lambda}$ evaluates to our sample mean. The exponential model with $\hat{\lambda}$ as our sample mean looks like a decent fit for our data because it appears to minimize the residuals from the superimposed 45° line.

(ii) $L(\lambda | y \not\in \beta) = \frac{1}{\lambda^n} e^{\left(-\sum_{i=1}^{n} y_i / \lambda\right)}$ joint distribution

Whereas $p(\lambda | \alpha) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{-(\alpha+1)} e^{-(\beta/\lambda)} & \text{for } \lambda > 0 \\ 0 & \text{o.w.} \end{cases}$

We can see that $\beta = \sum_{i=1}^{n} y_i$, and we must find a value for α such that the two equations are equivalent.

$$p(\lambda | \alpha, \beta) = \begin{cases} \frac{\left(\sum_{i=1}^{n} y_i\right)^\alpha}{\Gamma(\alpha)} \lambda^{-(\alpha+1)} e^{-\left(\sum_{i=1}^{n} y_i / \lambda\right)} & \text{for } \lambda > 0 \\ 0 & \text{o.w} \end{cases}$$

Our conjugate prior can also be rewritten as:

$C_+ \lambda^{-n} \exp^{-(S/\lambda)}$ where $S = \sum_{i=1}^{n} y_i$ (sum)

(iii) $p(\lambda | y \not\in \beta) = C_+ \left[p(\lambda | \alpha \beta)\right]\left[L(\lambda | y \not\in \beta)\right]$

$\left[\begin{matrix}\text{posterior dist} \\ \text{for } \lambda\end{matrix}\right] = \left[C_+ \lambda^{-(\alpha+1)} \exp\left(-\frac{\beta}{\lambda}\right)\right]\left[C_+ \lambda^{-n} \exp\left(\frac{-S}{\lambda}\right)\right]$

ignoring constants $= \lambda^{-(\alpha+1+n)} \exp^{-\left(\frac{\beta+S}{\lambda}\right)}$

Which is our inverse gamma function $\sim \Gamma^{-1}(\alpha+n, \beta+S)$

Where $S = \sum_{i=1}^{n} y_i \equiv \bar{y} n$

(iv) $E(\lambda) = \frac{\beta}{\alpha-1}$ for $\alpha > 1$ $\quad var(\lambda) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \alpha > 2$

prior mean $= \frac{\beta}{\alpha-1}$ $\quad (\alpha-1) = n_0$

Sample mean $= \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$

posterior mean $= \frac{\beta + \sum_{i=1}^{n} y_i}{\alpha-1+n} = \frac{\beta+s}{\alpha-1+n}$

Remember our parameters for our posterior I.G. distribution were

$$\Gamma^{-1}(\alpha+n, \beta+s)$$

Plugging those figures into our $E(\lambda)$ formula we get the same result for our posterior mean.

Seeing as we are adding $\alpha-1$ to n in our weighted average calculation, we can conclude that $\alpha-1$ is our prior effective sampling size.

(v) a) $\mu_0 = 4500$ $\quad \sigma_0 = 1,800$

$4500 = \frac{\beta}{\alpha-1}$ $\quad 3,240,000 = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$

$4500(\alpha-1) = \beta$

$\therefore \frac{4500^2(\alpha-1)^2}{(\alpha-1)^2(\alpha-2)} = 3,240,000$

$20,250,000 = 3,240,000\alpha - 6,480,000$

$\alpha = 8.25$

$\beta = 4500(8.25-1) = 32{,}625$

$\alpha = 8.25, \quad \beta = 32{,}625$

If $n_0 = \alpha - 1 = (8.25 - 1) = 7.25$,
then our sample size is about 7.

Given that in the EEE sample we had an $n = 14$, a sample size of 7 for our prior is quite large because it will influence our posterior distribution quite a bit given that it is half as large as our data set in terms of sample size.

b) Prior: $\Gamma^{-1}(8.25, 32625)$

where $\mu = 4{,}500, \quad \sigma = 1800$

Likelihood: $\Gamma^{-1}(13, \sum_{i=1}^{n} y_i)$

$\alpha = n - 1 = 13 \quad \beta = s = \sum_{i=1}^{n} y_i = 70612$

$\mu = \frac{\beta}{n} = \frac{70612}{14} = 5043 \qquad \sigma = \frac{70612}{\alpha - 1\sqrt{\alpha - 2}} = \frac{70612}{12\sqrt{11}} = 1774$

Posterior: $\Gamma^{-1}(\alpha + n, \beta + s)$ (α, β from prior)

$\alpha = 8.25 + 14 = 22.25$

$\beta = 32{,}625 + \sum_{i=1}^{n} y_i = 32{,}625 + 70{,}612 = 103{,}237$

$\mu = \frac{103{,}237}{22.25 - 1} = 4858$

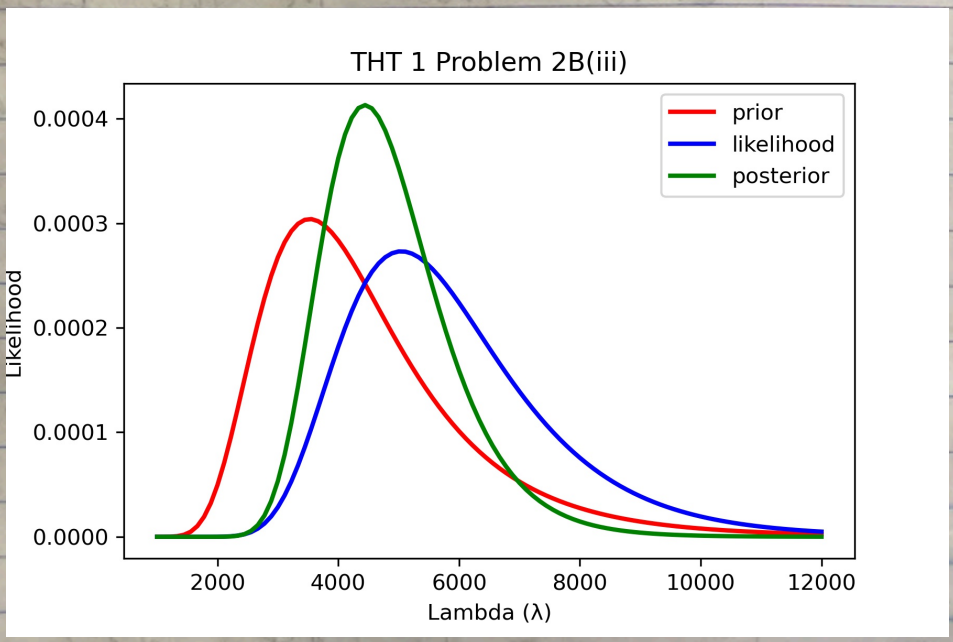
$$\sigma = \frac{\beta}{(\alpha-1)\sqrt{\alpha-2}} = \frac{103237}{21.25\sqrt{20.25}} = 1080$$

λ

	Prior	Likelihood	Posterior
Mean	4,500	5844	4,858
SD	1,800	1,774	1,080

c)

The posterior distribution is centered between the prior and sampling distributions. It has a tighter density curve because there is less uncertainty about lambda.



d) Posterior Frequentist Confidence Interval

$0.999 = Pr(4032.16 < \lambda < 5684.27)$

Posterior Bayesian Credible Interval

$0.999 = Pr(2511.78 < \lambda < 10423.19)$

These confidence intervals are so different because the Frequentist interval treats λ as fixed and 99.9% of confidence intervals would contain the parameter.

The Bayesian approach treats λ as an r.v. and gathers from its distribution the central interval in which there's a 99.9% probability

of containing λ.

Neither interval is "right" or "wrong", they just use different assumptions about the fixed/r.v. status of λ.