1. Suppose $X_1, X_2, \ldots, X_n$ are *iid* random variables with common pdf

$$f(x|\theta) = \theta x^{\theta-1},\ 0 < x < 1,\ \theta > 0$$

(a) Prove that $-\sum_{i=1}^{n} \ln(X_i)$ is a complete sufficient statistic for this family.

(b) Derive the probability distribution of $-\sum_{i=1}^{n} \ln(X_i)$.

(c) Find the $UMVUE$ for $\frac{1}{\theta}$. Does it attain the $CRLB$? Give a detailed reason why or why not. Show that the $UMVUE$ also happens to be the MLE of $\frac{1}{\theta}$.

(d) Find the MLE of θ and derive the asymptotic probability distribution of the MLE.

1. a)

$$f(\underset{\sim}{x}|\theta) = \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1} = \theta^n e^{(\theta-1)(\sum_i \ln x_i)}$$

$$= \frac{\theta^n}{e} \cdot e^{\theta(\sum_i \ln x_i)} \longrightarrow \circledast$$

re-write

$$f(\underset{\sim}{x}|\theta) = \frac{\theta^n}{e} \cdot e^{(-\theta)(-\sum_i \ln x_i)}$$

then this is exponential family with $w_1(\theta) = -\theta$

$t_1(x) = -\ln x$.

→ One parameter full exponential family.

Hence $-\sum \ln(x_i)$ complete sufficient.

Alternatively, Student's can keep $\circledast$ to argue exponential family with $\sum_i \ln x_i$ complete sufficient & since $-\sum \ln x_i$ is 1-to-1 function of $\sum_i \ln x_i$ it will be complete sufficient too.

1. b) Define, $Y_i = -\ln X_i$

$y = -\ln x \Rightarrow x = e^{-y} \Rightarrow \frac{dx}{dy} = -e^{-y}$, hence $|J| = e^{-y}$.

Thus

$$f_{Y_1}(y) = \theta(e^{-y})^{\theta-1} e^{-y}, \quad y > 0$$

$$= \theta e^{-y\theta}$$

$$= \frac{1}{\beta} e^{-y/\beta}, \quad y > 0, \text{ where } \theta = \frac{1}{\beta}$$

Hence $Y_i \sim$ Exponential(β) or Gamma$(1, \beta)$.

Then $-\sum_{i=1}^{n} \ln X_i = \sum_{i=1}^{n} Y_i \sim$ Gamma(n, β)

or Gamma$(n, \frac{1}{\theta})$

1.

c) Let us write

$$T = -\sum_{i=1}^{n} \ln x_i .$$

Since $T \sim \text{Gamma}\left(n, \frac{1}{\theta}\right)$ & T is complete sufficient for the family any function of T which is an unbiased estimator of $\frac{1}{\theta}$, should be the UMVUE of $\frac{1}{\theta}$.

From Gamma mean

$E_\theta(T) = \frac{n}{\theta}$, hence $E\left(\frac{T}{n}\right) = \frac{1}{\theta}$. Therefore UMVUE of $\frac{1}{\theta}$ is $\frac{T}{n}$.

Note that the likelihood function for θ is

$$L(\theta|t) = f(t|\theta) = \frac{\theta^n}{\Gamma n} t^{n-1} e^{-\theta t}$$

[students may do this in terms of $\beta = \frac{1}{\theta}$ too.]

$$\ln L(\theta|t) = \text{Const} + n \ln \theta - \theta t$$

Score function

$$\frac{d}{d\theta} \ln L(\theta|t)$$

$$= \left(\frac{n}{\theta} - t\right)$$

$$= -n\left(\frac{t}{n} - \frac{1}{\theta}\right) \longrightarrow \circledast\circledast$$

Note that the score function has form $a(\theta)[w(\underset{\sim}{x}) - \tau(\theta)]$ where $w(\underset{\sim}{x}) = \frac{t}{n}$, $\tau(\theta) = \frac{1}{\theta}$.

This form ensures attainment of CRLB by $w(\underset{\sim}{x}) = \frac{t}{n}$ for $\tau(\theta) = \frac{1}{\theta}$. So yes $\frac{T}{n}$ attains CRLB alternatively, students can seek to show $Var\left(\frac{T}{n}\right)$ is equal to CRLB in this case.

From Gamma properties

$$Var\left(\frac{T}{n}\right) = \frac{1}{n^2} Var(T) = \frac{1}{n^2} \cdot n \cdot \left(\frac{1}{\theta}\right)^2 = \frac{1}{n\theta^2}.$$

Since $\frac{T}{n}$ is an unbiased estimator of $\frac{1}{\theta}$,

$$CRLB = \frac{\left\{\frac{d}{d\theta} E_\theta\left(\frac{T}{n}\right)\right\}^2}{-E_\theta\left[\frac{d^2}{d\theta^2} \ln f(t|\theta)\right]}$$

$$= \frac{\left\{\frac{d}{d\theta}\left(\frac{1}{\theta}\right)\right\}^2}{-E\left(-\frac{n}{\theta^2}\right)}$$

$$= \frac{\left(-\frac{1}{\theta^2}\right)^2}{\frac{n}{\theta^2}} = \frac{\frac{1}{\theta^4}}{\frac{n}{\theta^2}} = \frac{1}{n\theta^2}$$

Thus $Var\left(\frac{T}{n}\right)$ attains CRLB.

Score function = 0 from ⊛⊛ gives

$$-n\left(\frac{t}{n}-\frac{1}{\theta}\right)=0$$

$$\Rightarrow \frac{t}{n}=\frac{1}{\theta}$$

Hence MLE of $\frac{1}{\theta}$ is $\frac{T}{n}$.

(d) From score function in (**),

$$-n\left(\frac{t}{n}-\frac{1}{\theta}\right)=0$$

$$\Rightarrow \theta = \frac{n}{t}$$

From that we can say MLE of θ is $\frac{n}{T}$.

Alternatively, MLE of $\frac{1}{\theta}$ is $\frac{T}{n}$ from (c), hence by property of MLE under continuous 1-to-1 transformation of θ, $\frac{n}{T}$ is the MLE

Note $\frac{T}{n}=\bar{Y}$ (where $Y_i = -\ln X_i \sim \text{Gamma}(1, \frac{1}{\theta})$) or Exponential$(\frac{1}{\theta})$

Hence by CLT

$$\sqrt{n}\left(\bar{Y}-\frac{1}{\theta}\right) \xrightarrow{d} N\left(0, \frac{1}{\theta^2}\right)$$

define $g(y)=\frac{1}{y}$, then $\frac{n}{T}=g(\bar{Y})$ & $g'(y) = -\frac{1}{y^2}$

By Delta method

$$\sqrt{n}\left[g(\bar{Y})-g(\tfrac{1}{\theta})\right] \xrightarrow{d} N\left(0, \{g'(\tfrac{1}{\theta})\}^2 \cdot \frac{1}{\theta^2}\right)$$

i.e

$$\sqrt{n}\left(\frac{1}{Y}-\theta\right) \xrightarrow{d} N\left(0, \left(-\frac{1}{1/\theta}\right)^2 \cdot \frac{1}{\theta^2}\right)$$

i.e

$$\sqrt{n}\left(\frac{1}{Y}-\theta\right) \xrightarrow{d} N\left(0, \theta^2\right)$$

Hence $\sqrt{n}\left(\frac{n}{T}-\theta\right) \xrightarrow{d} N(0,\theta^2)$ or $\frac{n}{T} \sim AN\left(\theta, \frac{\theta^2}{n}\right)$

Problem 2) a) Side-by-side boxplots show relatively symmetric distributions of salary for both males and females. Normal Q-Q plots show some evidence of non-normality. A two-sample t-test comparing the means is probably reasonable, but a distribution-free test might be preferred (or a t-test on some appropriate transformation of the data). In any case, the p-values are such that we do not conclude a difference in mean salaries between males and females.

```
t.test(yi ~ x1, var.equal=T)
    Two Sample t-test
t = 0.7101, df = 63, p-value = 0.4803
alternative hypothesis: true difference in means is not equal to 0
t.test(yi ~ x1)
    Welch Two Sample t-test
t = 0.7121, df = 58.886, p-value = 0.4792
alternative hypothesis: true difference in means is not equal to 0
wilcox.test(yi ~ x1)
    Wilcoxon rank sum test
W = 566, p-value = 0.5317
alternative hypothesis: true location shift is not equal to 0
```

b) An initial ordinary multiple linear regression fit shows some problems when the residual analysis is done. The residuals plotted against the fitted values (and plotted against x2) show a parabolic trend. There also appears to be non-constant error variance. We try a transformation; I tried including a quadratic term, x2^2. That fixed the nonlinearity, but the nonconstant error variance was still evident. I tried a natural log transformation of the response, which solved the problem nicely. Error variance now appears constant, and the normal error assumption seems reasonable as well (one slight outlier is noted from the Q-Q plot).

We examine the t-tests to look at the effect of each predictor on the response. X3 is nonsignificant, but x1 and x2 are significant in the chosen multiple regression model. We included an interaction term x1*x2 to determine whether the effect of x2 on the response depends on sex. This was not significant, so the interaction term was not used. The final model explains an extremely large percentage of the variation in monthly salary (99.5% based on R^2). It appears that females have a higher mean salary than males, for a given level of performance. And a higher performance rating yields a higher mean salary.

The model assumptions are that the random error terms are independent and normally distributed with mean zero and constant variance. The independence assumption is likely true since the data were taken cross-sectionally and from a random sample. The normality assumption is verified from a Q-Q plot of the residuals, and the constant variance assumption appears to hold based on the plot of residuals vs. fitted values. After the model is transformed, the functional form of the model appears correct, based on the residual analysis.

c) Based on the t-test about β_1, it appears that females have higher salaries than males, conditional on performance rating. This is a different conclusion than was made in part (a), when no difference by sex was found. It appears that accounting for performance rating alters whether sex affects mean salary. A pair of boxplots of performance rating, separate by sex, indicates that males seem to have slightly higher performance ratings than females. Perhaps this is the reason that the unconditional salary distributions of males and females do not appear to differ in terms of center.

d) We use the same model as in part (b), except we include as a predictor an indicator variable that is 1 if x3 > 100, and 0 otherwise. The t-test about the coefficient of this indicator is nonsignificant (two-sided P-value = 0.46), and in fact the estimated coefficient is negative, so the relevant one-sided P-value would be 1 – 0.23 = 0.77. So we conclude that there may be no difference between the mean salary of those who exceeded their personal budget and the mean salary of those who did not exceed their personal budget. A limitation of this inference is that only 5 of the 65 employees in the sample had x3 > 100, so we are essentially comparing a sample group of only 5 people to a sample group of 60 people, not ideal.

e) Answers for this one could vary, but a key is to recognize that formulas for the turning points can be found by taking the derivative of the mean response function with respect to M, setting this equal to zero, and using the quadratic formula to solve for the roots:

$$E(P) = \beta_0 + \beta_1 M + \beta_2 M^2 + \beta_3 M^3$$

Set $\frac{d}{dM} E(P) = \beta_1 + 2\beta_2 M + 3\beta_3 M^2 = 0$

Solving for M using quadratic formula:

$$M = \frac{-2\beta_2 \pm \sqrt{4\beta_2^2 - 4(3\beta_3)(\beta_1)}}{2(3\beta_3)}$$

$$\Rightarrow \quad M = \left(-\beta_2 - \sqrt{\beta_2^2 - 3\beta_1\beta_3}\right) / (3\beta_3)$$

$$\text{and } \left(-\beta_2 + \sqrt{\beta_2^2 - 3\beta_1\beta_3}\right) / (3\beta_3)$$

The point estimates of the turning points could be obtained by plugging the estimates in:
(-b2-sqrt(b2^2-3*b1*b3))/(3*b3) = 8389.268
(-b2+sqrt(b2^2-3*b1*b3))/(3*b3) = 12026.75

Then a good way to obtain approximate 95% familywise CIs for the two turning points is to generate many samples of "random error" values from a normal distribution with mean 0 and variance = MSE, and add these to the fitted values of the cubic regression fit. This creates kind of parametric bootstrap samples that use the normal error assumption. Then cubic regression could be fit on each bootstrap sample, and estimated turning points could obtained for each bootstrap sample, and then 90% bootstrap CIs for each turning point could be obtained by taking the 0.05 and 0.95 percentiles for each turning-point empirical distribution. By the Bonferroni method, this creates (conservative) 95% familywise intervals for the pair of turning points. My bootstrap intervals for an example run were: (8259.971, 8545.605) for TP1, and (11864.63, 12169.89) for TP2.

Suppose that

$$Y_i = m(x_i) + \varepsilon_i \ , \qquad i=1,\dots,n,$$

for fixed $x_i \in [0,1]$, and independent random variables ε_i s.t. $\mathbb{E}\,\varepsilon_i = 0$, $\mathbb{E}\,\varepsilon_i^2 = \sigma^2$, and where $m(\cdot)$ is a function from $[0,1]$ to $\mathbb{R}$.

Divide $[0,1]$ into L intervals of equal length

$$I_\ell = \left[\frac{\ell-1}{L}, \frac{\ell}{L}\right), \quad \ell = 1,\dots,L-1, \quad I_L = \left[\frac{L-1}{L}, 1\right]$$

and let $\mathbb{1}_\ell(x) = \begin{cases} 1 & \text{if } x \in I_\ell \\ 0 & \text{otherwise} \end{cases}$. Assume $n_\ell = \sum_{i=1}^n \mathbb{1}_\ell(x_i) > 0$ for $\ell = 1,\dots,L$.

For $\beta = (\beta_1,\dots,\beta_L)^T$, let $m_\beta(x) = \sum_{\ell=1}^L \beta_\ell \mathbb{1}_\ell(\cdot)$ be the function which takes the value β_ℓ over the interval I_ℓ.

Define $\hat{m}(\cdot) = m_{\hat{\beta}}(\cdot)$ where $\hat{\beta}$ is the minimizer of the least-squares criterion

$$\sum_{i=1}^n \left[Y_i - m_\beta(x_i) \right]^2 .$$

① Let $Y = (Y_1,\dots,Y_n)^T$, and find the design matrix Z such that

$$\sum_{i=1}^n \left[Y_i - m_\beta(x_i) \right]^2 = \| Y - Z\beta \|_2^2 \ , \qquad \text{where } \|a\|_2^2 = \sum_{i=1}^n a_i^2 .$$

<u>Answer</u>: Rewrite

$$\sum_{i=1}^n \left[Y_i - \sum_{\ell=1}^L \beta_\ell \mathbb{1}_\ell(x_i) \right]^2 \qquad Z_i = \left(\mathbb{1}_1(x_i), \dots, \mathbb{1}_L(x_i) \right)^T$$

$$= \sum_{i=1}^n \left[Y_i - Z_i^T \beta \right]^2$$

$$= \| Y - Z\beta \|_2^2 , \quad \text{with} \quad \underset{n \times L}{Z} = \begin{bmatrix} Z_1^T \\ \vdots \\ Z_n^T \end{bmatrix} = \begin{bmatrix} \mathbb{1}_1(x_1) & \mathbb{1}_2(x_1) & \cdots & \mathbb{1}_L(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{1}_1(x_n) & \mathbb{1}_2(x_n) & \cdots & \mathbb{1}_L(x_n) \end{bmatrix},$$

* Each row of Z has one nonzero entry equal to 1.

② Find $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_L)$.

Answer: We have

$Z^T Z$ is diagonal w/ entries $n_1, \dots, n_L$

$$\hat{\beta} = (Z^T Z)^{-1} Z^T Y = \begin{pmatrix} n_1 & & \\ & \ddots & \\ & & n_L \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} Y_i \mathbb{1}_1(x_i) \\ \vdots \\ \sum_{i=1}^{n} Y_i \mathbb{1}_L(x_i) \end{pmatrix} = \begin{bmatrix} \bar{Y}_1 \\ \vdots \\ \bar{Y}_L \end{bmatrix}$$

where $\bar{Y}_\ell = \frac{1}{n_\ell} \sum_{i=1}^{n} Y_i \mathbb{1}_\ell(x_i)$.

③ What is $\mathrm{Var}\left(m_{\hat{\beta}}(x)\right)$ for x in the interval I_ℓ?

Answer:

$$m_{\hat{\beta}}(x) = \bar{Y}_\ell \quad \text{for} \quad x \in I_\ell$$

$$\mathrm{Var}\left(\bar{Y}_\ell\right) = \frac{\sigma^2}{n_\ell}.$$

④ Suppose $m(\cdot)$ has the property that for any $x, y \in [0, 1]$,

$$|m(x) - m(y)| < C|x - y|,$$

for some $C > 0$.

What is an upper bound for

$$\mathbb{E}\left[m_{\hat{\beta}}(x) - m(x)\right]$$

for x in any interval I_ℓ?

Answer: $\mathbb{E}\ m_{\hat{\beta}}(x) - m(x) = \mathbb{E}\ \bar{Y}_\ell - m(x)$ for $x \in I_\ell$.

$$= \frac{\sum_{i=1}^{n} m(x_i)\, \mathbb{1}_\ell(X_i)}{n_\ell} - m(x)$$

$$= \frac{1}{n_\ell} \sum_{i=1}^{n} \mathbb{1}_\ell(x_i) \left[m(X_i) - m(x) \right]$$

$$\le \frac{1}{n_\ell} \sum_{i=1}^{n} \mathbb{1}_\ell(x_i)\, C\, |X_i - x|$$

$$\le \frac{C}{L}$$

(5) Increasing the number of intervals L will have what effect on the bias of $\hat{m}$?

On the variance of $\hat{m}$?

Answer: $L \uparrow \Rightarrow$ biss $\downarrow$

$L \uparrow \Rightarrow n_\ell \downarrow \Rightarrow$ variance $\uparrow$

Questions and Answers
Peña Question in 2017 PhD Quals Exam

Peña Question: While driving to work one morning, I got stopped in a traffic light at the intersection of Assembly Street and Gervais Street. I noticed that the vehicle in front of me had a broken brake light. I looked at the other vehicles that were also stopped (about 20 of them) and these other vehicles did not have broken brake lights (the police will usually give you a ticket if your brake lights are broken, so the event of a car having broken brake lights is a rare event). I got to wondering: *What is the percentage of all vehicles being driven in Metropolitan Columbia have broken brake lights?* You *are* the statistician that is consulted about this problem.

(a) Describe a statistical sampling plan that could be done within a reasonable amount of time at a reasonable cost to gather relevant data to answer the primary question of inferring about the proportion, denoted by θ, of all vehicles being driven in MetroColumbia that have broken brake lights. Explain why your design is appropriate and describe the type of sample data that you will obtain from this study. In particular will your study have a fixed sample size, or will it have a random sample size?

Solutions: There are several possible sampling plans for this problem, which could be justified to be appropriate based on their arguments.

It could be *fixed sample size plan*, with (not viable in practice) or without replacement, and this could just be a simple random sample or it could be a stratified random sample. Most likely the student will use a simple random sample. An important detail to look for in their answer is how the sampling process will be implemented in practice and how they will decide on their sample size. They might mention listing all cars registered in SC and sampling from them, but this is not the most appropriate since we want those driven in Metro Columbia. Most appropriate might be a scheme where cars passing through sampled traffic intersections in Metro Columbia are sampled and observed for broken brake lights. But there are many possible ideas that could be put forward and so long as they could justify their scheme in a reasonable manner, that will be fine.

A second possible sampling plan is where a fixed integer k is specified, say, 20, and sampling continues until k vehicles with broken brake lights are observed. Such random sample size sampling scheme may be most appropriate in this case since the event of a broken brake light is rare as indicated in the statement of the problem. If their answer mentions the rarity of the event and uses a random sample size sampling scheme, this is good since this indicates critical thinking by the student.

(b) Based on the data that you will obtain from your study, describe and fully justify your procedure for performing inference (estimation and constructing a confidence interval) about θ. You should describe the appropriate statistical model that you will be postulating and must justify why such a model is reasonable. You should describe the estimator that you will use and justify why such an estimator will be good. For instance, will your estimator have desirable properties and what are these desirable properties? You should also describe how you will obtain a measure of the degree of precision of your estimator.

Solutions: The answer to this depends on the sampling scheme stated in item (a). If a fixed sample size scheme was used, and the student either mentions that even though it is without replacement but that the number of cars in the population of interest will be large so it could be assumed that it is sampling with replacement, hence a Bernoulli or binomial model could be used, then this will be fine. If a hypergeometric model is mentioned, then this is even better as this recognizes that with sampling without replacement, the independence needed under the binomial model may break down. The estimator depends on whether a binomial model or a hypergeometric model is assumed as the statistical model. In both cases, the proportion of vehicles with damaged brake lights in the observed sample will be a reasonable estimator as it will be the MM or ML estimator. That is, a reasonable estimator will be X/n where X is the number with broken brake lights among the n sampled vehicles.

If a random sample size scheme is used, then the appropriate statistical model will be a negative binomial model, provided that either the student mentions sampling with replacement or that independence could be assumed since the population size is large. The MM estimator will then be k divided by the number of sampled cars by noting that the expected value of the number of cars sampled will be k/θ. In fact, it is also the ML estimator of θ by noting that the likelihood function is proportional to $\theta^k(1-\theta)^{N-k}$ where N is the total sample size.

Estimates of the standard errors of the estimate will depend on which sampling scheme and statistical model is assumed above, with the SE estimate obtained via a plug-in procedure. Of course the student could use asymptotic arguments to obtain their standard error estimate, and this will be fine. The important aspect to look for is whether they could determine how the standard error of their estimator will depend on the sampling scheme and the statistical model assumed.

(c) Based on past information about vehicles in Metropolitan Columbia, a prior distribution about θ is given by a beta distribution with parameters $(\alpha, \beta) = (2, 98)$, so that the prior density function is

$$\pi(\theta) = \frac{1}{B(2,98)}\theta^{2-1}(1-\theta)^{98-1}I\{0 < \theta < 1\}.$$

If you are given this prior information, what will be your Bayes estimator of θ based on squared-error loss function?

Solutions: For the binomial and negative binomial statistical models, the posterior distribution of θ will be proportional to

$$\theta^{2+x-1}(1-\theta)^{98+n-x-1}$$

where x is the number with broken brake lights and n is the number of vehicles sampled. The Bayes estimator of θ based on squared-error loss will then be the posterior mean, which will be

$$\hat{\theta} = \frac{X+2}{100+n}.$$

If a hypergeometric statistical model is assumed, then this leads to a more complicated Bayes estimate since the prior will not anymore be a conjugate prior. If a student decides this route and mentions using numerical tools to obtain the posterior mean of θ, then that shows excellent understanding and should be given high credit.

(d) Based on your sampling design, could your estimates obtained in (b) and also in (c) be equal to zero? If you get an estimate of zero, will this be a sensible or reasonable estimate?

Solutions: Under the binomial statistical model, and also the hypergeometric statistical model, the estimate could end up becoming zero. For the negative binomial statistical model, it will never be equal to zero. The student might mention in passing that in this scheme the needed sample size could be quite large and that is an astute observation and should be viewed in very positive light. The Bayes estimate will never be equal to zero.

An estimate of zero could still be argued as reasonable if the student mentions that this does not truly imply that θ is actually equal to zero, but that this indicates that θ has truly a small value. The student may also state that this will be unreasonable since based on the statement of the problem, at least one car was seen to have a broken brake light. So it depends on how they will argue about the reasonableness of such a zero estimate.

(e) A certain Professor X, who is *not* so knowledgeable about the intricacies of statistical modeling and inference, insisted that the best sampling design for this study is to observe 500 randomly chosen cars in Metropolitan Columbia. Upon making his observations (of course, with the help of his willing students), he found that none of the 500 cars that were observed have broken brake lights. However, he still claims that a conservative 95% confidence interval for θ based on the observed data is given by $[0, 3/500] = [0, .006]$. Is he justified in his claim? Justify your answer.

Solutions: This is a fixed sampling plan that was used by Professor X. Under a binomial model, if Y is the number of vehicles with broken brake lights among the 500 sampled vehicles, then Y has a binomial distribution with parameters $n = 500$ and θ. Under this model, the probability of $Y = 0$ is going to be

$$\Pr_\theta(Y = 0) = (1 - \theta)^{500}.$$

Thus, the set of θ-values such that $\Pr_\theta(Y = 0) = (1 - \theta)^{500} \geq 0.05$ are those which are less than or equal to $1 - (.05)^{1/500} = 0.000597$. The value of $3/500 = .006$, whereas $4/500 = .008$, hence the interval $[0, 3/500]$ is indeed a conservative 95% CI for θ when $Y = 0$ is observed. Thus, Professor X's claim is justifiable.

(f) Using the sample data obtained by Professor X and the prior distribution of θ in item (c), what would be a 95% Bayesian credible interval for θ?

Solutions: From the answer in (c) and the data in (e), the posterior distribution of θ is going to be a beta distribution with parameters $(2+y, 98+n-y) = (2, 598)$. Limits of a 95% credible interval for θ are therefore

$$\texttt{LOWER LIMIT} = qbeta(.025, 2, 598) = 0.000405;$$

$$\texttt{UPPER LIMIT} = qbeta(.975, 2, 598) = 0.009266.$$

REMARK on Grading: I will be assigning a total of 10 points for each of these 6 items for a total of 60 points. Then I'll just convert to a percentage score. Graders may use a different scheme but the score for the whole problem is one percentage score, e.g., 75%.

Problem 5) a) The likely reason is that there is some differences in growing conditions (field condition? sunlight/shade amount?) that will affect the yield. A way to account for this variation is to make, say, the regions of the field (or whatever affects the growing conditions) the blocks. The "random block effects" assumption implies that the researcher believes the growing conditions used in this experiment are a random selection from some large potential population of growing conditions.

b) $Y_{ij} = \mu_{..} + \rho_i + \tau_j + \varepsilon_{ij}$, $i = 1,\ldots,4$, $j=1,\ldots 13$.
ρ_i = effect of i-th block
τ_j = effect of j-th treatment
Here, $\mu_{..}$ is a constant, $\rho_i \sim$ independent $N(0, \sigma_\rho^2)$, $\Sigma\tau_j = 0$, $\varepsilon_{ij} \sim$ independent $N(0, \sigma^2)$, and ε_{ij} are independent of ρ_i.

c)

Source	df	F*
Blocks	3	MSB/MSE (a.k.a. MSB/MSBL.TR)
Treatments	12	MSTR/MSE (a.k.a. MSTR/MSBL.TR)
Error (a.k.a. Blk×Trt)	36	
Total	51	

d) We would simultaneously test these 12 contrasts (where the "new" variety is labeled variety 13, say): H_0: $\mu_1 - \mu_{13} = 0$ vs. H_a: $\mu_1 - \mu_{13} < 0$, H_0: $\mu_2 - \mu_{13} = 0$ vs. H_a: $\mu_2 - \mu_{13} < 0$, …, H_0: $\mu_{12} - \mu_{13} = 0$ vs. H_a: $\mu_{12} - \mu_{13} < 0$. A good approach might be to use a Scheffe multiple comparisons procedure, since this is designed to test a large number of contrasts. (Using Dunnett's procedure, which tests every treatment against one control, would also be an excellent approach.) The Bonferroni method would be possible, but not ideal here since there are a large number of contrasts, and the Bonferroni method would be quite conservative. The Tukey method would not be ideal since we are not comparing all pairs of treatment means (not even close to all pairs, in fact).

e) $$\text{var}(\bar{Y}_{\cdot j}) = \text{var}\left(\frac{1}{4}\sum_{i=1}^{4} Y_{ij}\right) = \frac{1}{16}\sum_{i=1}^{4}\text{var}(Y_{ij})$$

$$= \frac{1}{4}(\sigma^2 + \sigma_\rho^2) \Rightarrow sd(\bar{Y}_{\cdot j}) = \frac{1}{2}\sqrt{\sigma^2 + \sigma_\rho^2}$$

$$\text{var}(\bar{Y}_{i\cdot}) = \text{var}\left(\frac{1}{13}\sum_{j=1}^{13} Y_{ij}\right) = \frac{1}{13^2}\,\text{var}\left(\sum_{j=1}^{13} Y_{ij}\right)$$

$$= \frac{1}{13^2}\left[13(\sigma^2 + \sigma_\rho^2) + 2\binom{13}{2}\text{cov}(Y_{ij}, Y_{ij'})\right]$$

$$= \frac{1}{13}(\sigma^2 + \sigma_\rho^2) + \frac{156}{169}\sigma_\rho^2 = \frac{1}{13}\sigma^2 + \sigma_\rho^2$$

$$\Rightarrow sd(\bar{Y}_{i\cdot}) = \sqrt{\frac{1}{13}(\sigma^2 + \sigma_\rho^2) + \frac{156}{169}\sigma_\rho^2}$$

$$\text{or} \quad = \sqrt{\frac{1}{13}\sigma^2 + \sigma_\rho^2}$$

f) $P\left[\frac{1}{4}\sum_{k=1}^{4} Y_{11k} \geq 1.3(\mu_{..}+\tau_3)\right]$ (*)

since $E(\bar{Y}_{13.}) = \mu_{..} + \tau_3$.

And since $\frac{1}{4}\sum_{k=1}^{4} Y_{11k} \sim N\left(\mu_{..}+\tau_1, \sigma_p^2 + \frac{\sigma^2}{4}\right)$,

$$(*) = P\left[Z \geq \frac{1.3\mu_{..} + 1.3\tau_3 - \mu_{..} - \tau_1}{\sqrt{\sigma_p^2 + \frac{\sigma^2}{4}}}\right]$$

$$= 1 - \Phi\left(\frac{0.3\mu_{..} + 1.3\tau_3 - \tau_1}{\sqrt{\sigma_p^2 + \frac{\sigma^2}{4}}}\right)$$

6. Suppose $X_1, X_2, \ldots, X_{30}$ is a random sample from a $Gamma(2, \theta)$ distribution.

(a) Derive the size 0.05, UMP test for $H_0 : \theta \leq 1$ $vs.$ $H_1 : \theta > 1$ and derive the power function of the test.

(b) Consider the testing problem $H_0 : \theta = 1$ $vs.$ $H_1 : \theta \neq 1$. Argue that an UMP test does not exist in this case. Derive the exact size 0.05 Likelihood Ratio Test (LRT) for this problem.

(c) In the same graph, plot the power function of the test in part (b) along with the power function in part (a) and comment on the plot.

(d) Give a 95% confidence interval for θ using a random sample of size 30, viz. $X_1, X_2, \ldots, X_{30}$.

(e) Suppose the true value of θ is 1. Generate 30 random observations from $Gamma(2, 1)$ distribution and construct the confidence interval derived in (d) from this generated sample observations. Repeat this 10 times to generate 10 different confidence intervals. What is the observed coverage probability (proportion of these 10 constructed intervals that actually covers the true value $\theta = 1$)? Provide the code used for this simulation.
Comment on why the observed coverage probability may not be exactly equal to the confidence coefficient 95% used in part (d).

2.a)

The joint distribution is given by

$$f(\underset{\sim}{x}|\theta) = \left(\frac{1}{\Gamma 2\,\theta^{2}}\right)^{30} \left(\prod_{i=1}^{30} x_i^{2-1}\right) e^{-\sum_{i=1}^{30} x_i/\theta}$$

→ One parameter full exponential family,

$T = \sum_{i=1}^{30} X_i$ Complete, sufficient for the family.

Also,

T is a sum of iid Gamma, so $T \sim \text{Gamma}(60, \theta)$

pdf

$$g_{T|\theta}(t) = \frac{1}{\Gamma 60\; \theta^{60}}\, t^{60-1}\, e^{-t/\theta}, \quad t > 0$$

$$= \frac{1}{59!} \cdot \frac{1}{\theta^{60}} \cdot t^{59} \cdot e^{\left(-\frac{1}{\theta}\right)t}$$

$$= c(\theta) \cdot h(t)\, e^{\omega(\theta)\, t}$$

where, $\omega(\theta) = -\frac{1}{\theta}$ which is increasing in θ.

Hence T has an MLR.

Then the UMP test for $H_0: \theta \leq 1$ vs. $H_1\; \theta > 1$, by Karlin Rubin theorem, is

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & \text{if } T > c \\ 0 & \text{o.w} \end{cases}$$

where c is such that $\sup_{\theta \in \Theta_0} E_\theta[\phi(\underset{\sim}{x})] = 0.05$

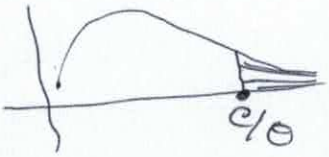
i.e $\sup_{\theta \leq 1} E_\theta[\phi(\underset{\sim}{x})] = 0.05$

Now,

$$E_\theta[\phi(\underset{\sim}{x})]$$
$$= P_\theta(T > c)$$
$$= P\left(\frac{T}{\theta} > \frac{c}{\theta}\right)$$
$$= P\left(T^* > \frac{c}{\theta}\right)$$

$\left[T \sim \text{Gamma}(60, \theta) \right.$
$\left. \text{Hence } T^* = \frac{T}{\theta} \sim \text{Gamma}(60, 1) \right]$

$$= 1 - F\left(\frac{c}{\theta}\right)$$ where F is the CDF of Gamma(60, 1)



Since $E_\theta[\phi(\underset{\sim}{x})] \uparrow \theta$

$$\sup_{\theta \in \Theta_0} E_\theta[\phi(\underset{\sim}{x})] = E_{\theta=1}[\phi(\underset{\sim}{x})]$$
$$= 1 - F(c)$$

~~zote~~

For a size 0.05 test

$$1 - F(c) = 0.05$$
hence $F(c) = 0.95$
$$c = \text{qgamma}(0.95, 60, 1) = \underline{73.28868}$$

Thus size 0.05 UMP test is

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{30} x_i > 73.28868 \\ 0 & \text{otherwise} \end{cases}$$

Power function

$$\beta(\theta) = E_\theta[\phi(\underline{x})]$$

$$= P_\theta(T > 73.28368)$$

$$= 1 - F_{T|\theta}(73.28368)$$

[Here $T|\theta \sim$ Gamma(60,

(in R $\beta(\theta) = 1 -$ pgamma(73.28368, 60, scale = θ))

——×——

Note:

Students may choose to solve for c just doing $P_{\theta=1}(T>c) = 0.05$ without exclusively showing $\sup_{\theta \in \Theta_0} E_\theta[\phi(\underline{x})]$ is achieved at $\theta = 1$. An MLR will give that.

But showing $E_\theta(\phi(\underline{x}))$ is increasing in θ & therefore size is achieved at $\theta = 1$, should earn bonus points.

b) Since T has MLR, the UMP test of size 0·05 for $H_0: \theta \geq 1$ vs. $H_1: \theta < 1$ is given by

$$\phi^*(\underset{\sim}{x}) = \begin{cases} 1 & \text{if } T < c^* \\ 0 & \text{otherwise.} \end{cases}$$

where c^* is chosen so that

$$\sup_{\theta \geq 1} E_\theta(\phi^*(\underset{\sim}{X})) = 0{\cdot}05$$

$$\Rightarrow \sup_{\theta \geq 1} P_\theta(T < c^*) = 0{\cdot}05$$

$$\Rightarrow P_{\theta=1}(T < c^*) = 0{\cdot}05$$

$$\Rightarrow c^* = \text{qgamma}(0.05, 60, 1) = 47{\cdot}85232$$

Thus $$\phi^*(\underset{\sim}{x}) = \begin{cases} 1 & \text{if } T < 47{\cdot}85232 \\ 0 & \text{otherwise} \end{cases}$$

Note that ϕ from part a) gives higest power possible in $\theta > 1$ region.
& ϕ^* here gives highest power possible in $\theta < 1$ region.
(Since both are UMP).

Now, if a UMP of same size (0.05) exists for

$$H_0: \theta = 1 \text{ vs. } H_1: \theta \neq 1,$$

by the uniqueness of UMP tests, that particular test has to be same as both ϕ & ϕ^*.
But ϕ & ϕ^* have complementary rejection region so no single test can be same as both ϕ & ϕ^* simulteneously.
Hence UMP test for $H_0: \theta = 1$ vs. $H_1: \theta \neq 1$ does not exist.

T is sufficient so we can look at the likelihood function based on T.

$$L(\theta|t) = \frac{1}{59!\,\theta^{60}} t^{59} e^{-t/\theta}$$

$$\Rightarrow \ln L(\theta|t) = \text{Const} - 60\ln\theta - t/\theta$$

$$\frac{d}{d\theta}\ln L(\theta|t) = 0 \Rightarrow \frac{-60}{\theta} + \frac{t}{\theta^2} = 0$$

$$\Rightarrow \hat{\theta} = t/60$$

MLE of θ is $T/60$.

LR: $$\lambda = \frac{L(\theta=1|t)}{L(\theta=\hat{\theta}|t)}$$

$$= \frac{\frac{1}{59!} t^{59} e^{-t}}{\frac{1}{59!}\frac{1}{(t/60)^{59}} t^{59} e^{-60}}$$

$$= \text{Const}\ t^{60} e^{-t}$$

LRT: $$\phi(\underset{\sim}{x}) = \begin{cases} 1 & \text{if } \lambda < c \\ 0 & \text{otherwise.} \end{cases}$$

$$\Rightarrow \phi(\underset{\sim}{x}) = \begin{cases} 1 & \text{if } t^{60}e^{-t} < c \\ 0 & \text{otherwise} \end{cases}$$

let us write:

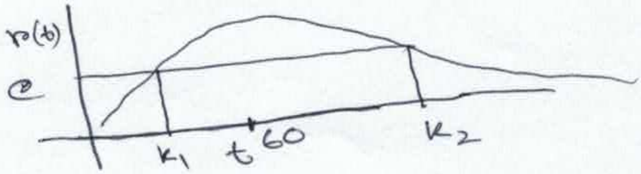
$$r(t) = t^{60}e^{-t}$$

$$r'(t) = 60\,t^{59}e^{-t} - t^{60}e^{-t}$$

$$= t^{59}e^{-t}(60-t)$$

$r'(t)$ -ve for $t<60$.

$r'(t)$ +ve for $t>60$ & $r'(t)$ -ve for $t<60$.



Hence LRT is given by

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & \text{if } T < k_1 \text{ or } T > k_2 \\ 0 & \text{o.w} \end{cases}$$

we choose k_1 & k_2 s.t

$$E_{\theta \neq 1}(\phi(\underset{\sim}{x})) = 0.05$$

One choice

$$P_{\theta=1}(T < k_1) = 0.025$$

$$P_{\theta=1}(T > k_2) = 0.025$$

$k_1 =$ qgamma(0.025, 60, scale=1) = 45.78632.

$k_2 =$ qgamma(0.975, 60, scale=1) = 76.1057.

Thus size 0.05 LRT is

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & \text{if } T < 45.78632 \text{ or } T > 76.1057 \\ 0 & \text{otherwise.} \end{cases}$$

c) The power function in part a) was

$$\beta(\theta) = 1 - F_{T|\theta}(73.28368)$$

in R

$$\beta(\theta) = 1 - \text{pgamma}(73.28368,\ 60,\ \text{scale} = \theta)$$

power function for LRT

$$\beta_L(\theta) = E_\theta\left(\phi_{LRT}(\underset{\sim}{x})\right)$$

$$= P_\theta(t < 45.78632) + P_\theta(T > 76.1057)$$

$$= F_{T|\theta}(45.78632) + 1 - F_{T|\theta}(76.1057)$$

(where $T \sim \text{Gamma}(60, \theta)$)

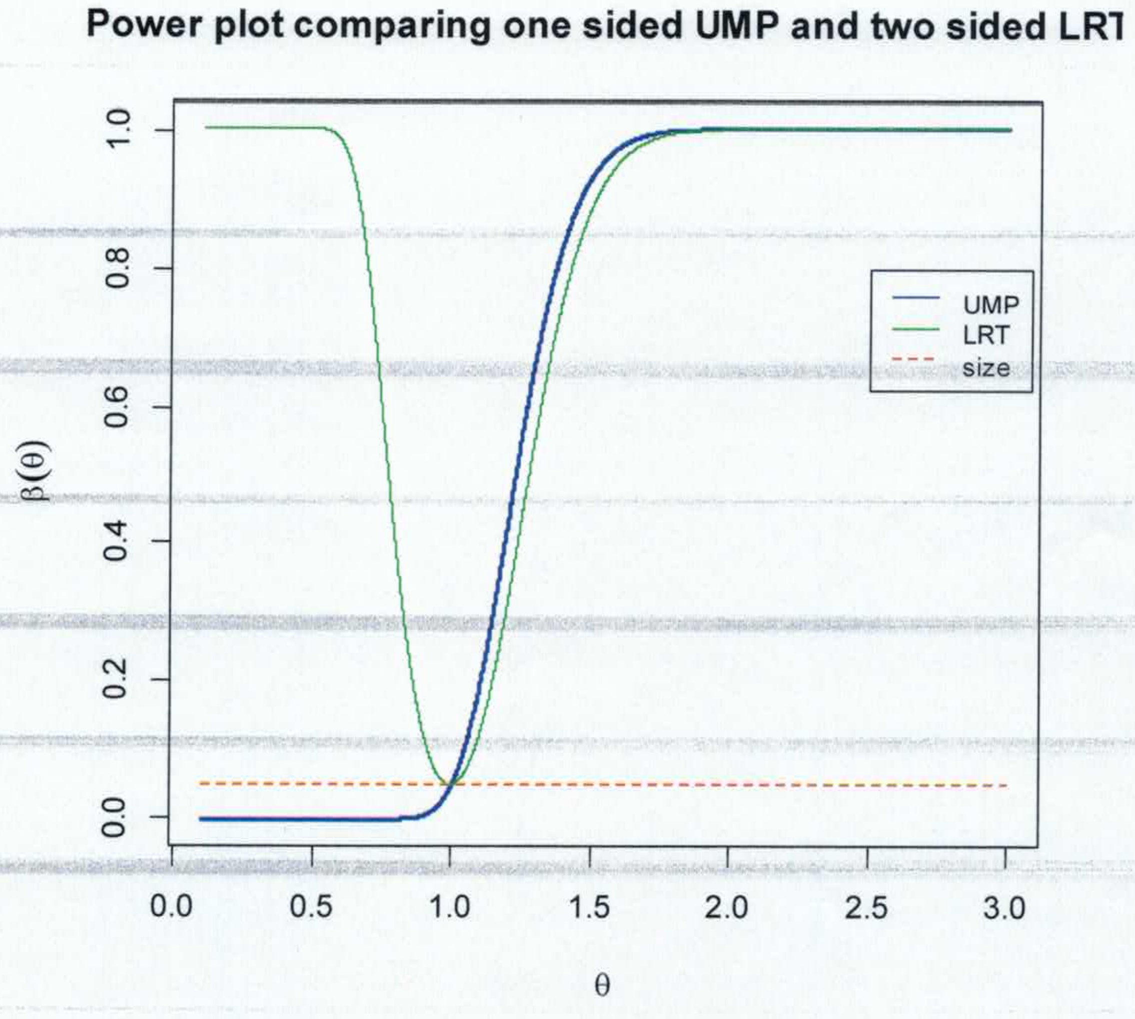
in R

$$\beta_L(\theta) = \text{pgamma}(45.78632,\ 60,\ \text{scale} = \theta)$$
$$+ 1 - \text{pgamma}(76.1057,\ 60,\ \text{scale} = \theta)$$

2. c) Both LRT and UMP attains size 0.05 at θ=1.

Since the UMP test is one sided and most powerful for $H_0: \theta \leq 1 \; vs \; H_1: \theta > 1$, it shows more (highest possible) power than the LRT in $\theta > 1$ region.

But the LRT, being a test for $H_0: \theta = 1 \; vs \; H_1: \theta \neq 1$, will show power much greater than the UMP above in $\theta < 1$ region.



R code:

```
theta=seq(0.1, 3, by=0.001)
n=length(theta)
x=beta1=beta2=matrix(0,nrow=n,ncol=1)
x=matrix(0.05, nrow=n,ncol=1)
```

```
for(i in 1:n)
{beta1[i]=1-pgamma(73.28368,60,scale=theta[i])}
for(i in 1:n)
{beta2[i]=pgamma(45.78632,60,scale=theta[i])+1-pgamma(76.1057,60,scale=theta[i])}
plot(theta,beta1,type="l",col="blue", xlab=expression(theta), ylab=expression(beta(theta)),
lwd=2, main="Power plot comparing one sided UMP and two sided LRT")
lines(theta,beta2,type="l",col="green", lwd=1.7)
lines(theta,x,type="l", lty=2, col="red")
legend(2.5, 0.8, legend=c("UMP", "LRT", "size"),col=c("blue", "green", "red"), lty=c(1,1,2),
cex=0.8)
```

d) There are number of ways of doing this.

(i) $T \sim$ Gamma $(60, \theta)$

Hence $\frac{T}{\theta} \sim$ Gamma $(60, 1)$

If $k_1 =$ qgamma$(0.025, 60, 1) = 45.78632$

$k_2 =$ qgamma$(0.975, 60, \text{scale}=1) = 76.1057$

Then

$$P_\theta\left(45.78632 < \frac{T}{\theta} < 76.1057\right) = .95$$

Hence a 95% C.I for θ is

$$\left[\frac{T}{76.1057}, \frac{T}{45.78632}\right].$$

(ii) Pivoting the cdf of T.

Suppose t_0 is an observed value of T, then choose θ_L & θ_U st

$$\int_0^{t_0} f_{T|\theta}(t)\,dt = .025, \quad F_{T|\theta_U}(t_0) = .025$$

$$\int_0^{t_0} f_{T|\theta}(t)\,dt = 0.975, \quad F_{T|\theta_L}(t_0) = .975$$

(Practically will give same C.I as (i) but students may leave it in this form)

(iii) Large Sample C.I

$$\frac{T}{n} = \bar{X}, \quad \left(\text{here } \frac{T}{30} = \bar{X}\right), \qquad E(X) = 2\theta, \quad Var(X) = 2\theta^2$$

$$\sqrt{n}\left(\bar{X} - 2\theta\right) \xrightarrow{d} N\left(0, 2\theta^2\right)$$

$$\frac{\bar{X} - 2\theta}{(\sqrt{2}\theta)/\sqrt{n}} \sim AN(0,1)$$

$$P_\theta\left(-1.96 < \frac{\bar{X} - 2\theta}{\sqrt{2}\theta/\sqrt{n}} < 1.96\right) \approx .95 \longrightarrow (*)$$

If we estimate Var(X) by sample variance S^2, then by Slutsky's theorem

$$P_\theta\left(-1.96 < \frac{\bar{X} - 2\theta}{s/\sqrt{n}} < 1.96\right) \approx .95$$

Asymptotic C.I

$$\left[\frac{\bar{X} - 1.96\, s/\sqrt{n}}{2}, \; \frac{\bar{X} + 1.96\, s/\sqrt{n}}{2}\right]$$

For this data,

$$\left[\frac{\bar{X} - 1.96\, s/\sqrt{30}}{2}, \; \frac{\bar{X} + 1.96\, s/\sqrt{30}}{2}\right].$$

e) The answer will depend om what confidence interval a student will construct.

Note that with 10 intervals the observed coverage probability (percentage of intervals that actually contain θ=1) can be 70% (7 out of 10), 80% (8 out of 10), etc upto 100% (10 out of 10). But never exactly 95%.

Following is an example of 10 generated intervals of form $\left[\frac{T}{76.1057}\ ,\ \frac{T}{45.78632}\right]$

Rcode:

```
m=matrix(0,ncol=2, nrow=10)
T=matrix(sum(rgamma(30,2,1)),nrow=10,ncol=1)
for (i in 1:10)
{T[i]=sum(rgamma(30,2,1))}
for ( i in 1:10)
{m[i,1]=T[i]/76.1057
m[i,2]=T[i]/45.78632}
m
```

R output:

```
      [,1]    [,2]
[1,] 0.7439902 1.236655
[2,] 0.9381310 1.559355
[3,] 0.7579994 1.259941
[4,] 0.9216290 1.531925
[5,] 0.8541957 1.419838
[6,] 0.7967779 1.324399
[7,] 0.6245257 1.038082
[8,] 0.7886819 1.310942
[9,] 0.7554194 1.255653
[10,] 1.0237514 1.701672
```

Each row given a confidence interval for θ. Note that here 9 out of 10 intervals captured θ=1. Hence observed coverage probability 90%.

Comment: The confidence coefficient 95% is associated with the sampling distribution of T, as
$P\left(\theta \in \left[\frac{T}{76.1057}\ ,\ \frac{T}{45.78632}\right]\right) = 0.95.$
Here the observed coverage probability is observed as a relative frequency of capture out of only 10 intervals from 10 realizations of T. If we keep on constructing the intervals and try to see the relative frequency of capture out of n intervals, the relative frequency will converge to true probability 95% as $n \to \infty$.
In other words, if the number of generated confidence interval is large (much larger than mere 10) we will see observed coverage percentage getting close to nominal value 95%.