Statistical Theory Final Project

Yisrael Haber, Hadar Kaner

September 2, 2021

Contents

1 First Problem

For $\alpha > 0$ define the gamma function -

$$\Gamma(\alpha) := \int_0^\infty t^{\alpha-1}e^{-t}dt$$

1. Show that for all $\alpha > 0$ the integral converges and infer that the following function is a distribution -

$$f(t) := \frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)}$$

(Correction - f isn't exactly a distribution but $f'(t) := \frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)} \cdot \mathbb{1}_{\{t\geq 0\}}$is. In the solution we will talk about f' as f.)

2. Show that for all $\alpha > 0$ we have that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, find a convenient expression for $\Gamma(n)$ for $n \in \mathbb{N}$.

3. Use a variable substitution on the distribution in (1) to get the gammma distribution

$$X \sim \Gamma(\alpha,\beta) \iff f_X(x) := \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}, & x > 0 \\ 0, & \text{Otherwise} \end{cases}$$

4. Show that if $X \sim \Gamma(\alpha,\beta)$ then show that $\mathbb{E}[X] = \alpha\beta$ and $\text{Var}(x) = \alpha\beta^2$. Find a general expression for $\mathbb{E}[X^c]$ for any $c > 0$.

5. Is the gamma distribution in the exponential family of distributions?

6. Let $X_1, \ldots, X_n$ samples from the distribution $\Gamma(\alpha,\beta)$. Find a sufficient statistic for (α,β).

7. Let $X_1, \ldots, X_n$ samples from the distribution $\Gamma(\alpha,\beta)$. Assume α is known and show that $\sum_i X_i$ is a minimal sufficient statistic.

8. Assume X is normal standard distribution, And take $Y = X^2$. Show that $Y \sim \Gamma(\frac{1}{2}, 2)$.

Solution:

1. Notice that we can use the following claim

> **Claim 1.** For every $\beta \in \mathbb{R}$, there is $t_0 > 0$ such that for every $t > t_0$ we have that $t^\beta < e^t$
>
> *Proof.* If $\beta < 0$ then this is obviously true since the term on the left tends to 0 and the term on the right tends to infinity and so this is true essentially directly from the definition of a limit. This can be seen by picking some positive value. Eventually the term on the right will remain over that constant, and the term on the left will remain under this constant and you will get this inequality as a consequence.
>
> Otherwise take $\beta \geq n \in \mathbb{N}$ notice that we can consider the fraction $\frac{e^t}{t^n}$. Both the top and the bottom tend to infinity and so we can apply L'hopital's rule, we can do this recursively until the bottom is a constant. The top will continue to stay the same throughout this process. This means the fraction will tend to infinity, and so also the original fraction will tend to infinity. This means eventually $\frac{t^n}{e^t} < 1$, using simple arithmetic we get our original claim. $(t^\beta < t^n < e^t)$ ■

Now we can notice that this means that for $0 < \alpha \in \mathbb{R}$ we have that there is $t_0 \in \mathbb{R}$ such that for all $t > t_0$ we have that $t^{\alpha+1} < e^t$. Meaning $t^{\alpha-1}e^{-t} < t^{-2}$. We can now use this to get -

$$\Gamma(\alpha) = \int_0^{t_0} t^{\alpha-1}e^{-t}dt + \int_{t_0}^{\infty} t^{\alpha-1}e^{-t}dt \leq \int_0^{t_0} t^{\alpha-1}dt + \int_{t_0}^{\infty} \frac{1}{t^2}dt =$$

$$= \left[\frac{t^\alpha}{\alpha}\right]_0^{t_0} + \left[-t^{-1}\right]_{t_0}^{\infty} = \left(\frac{t_0^\alpha}{\alpha} - 0\right) + \left(\frac{1}{t_0} - 0\right) = \frac{t_0^\alpha}{\alpha} + \frac{1}{t_0} < \infty$$

And so our integral converges. Remember that a distribution is a non-negative function whose total integral is 1. Obviously we have that the above defined (corrected) f is a distribution - the integral is positive since the integrand is positive, and we normalized the integral to give us a total integral of 1. Alltogether the integral convergesand f is a distribution. ■

2. We will use integration by parts

$$\Gamma(\alpha+1) = \int_0^{\infty} t^\alpha e^{-t}dt$$

$$\text{Integration By Parts } = \begin{bmatrix} u = t^\alpha, & v' = e^{-t} \\ u' = \alpha t^{\alpha-1}, & v = -e^{-t} \end{bmatrix}$$

$$\begin{aligned} \Gamma(\alpha+1) &= \int_0^{\infty} t^\alpha e^{-t}dt \\ &= \left[-t^\alpha e^{-t}\right]_0^{\infty} - \left(\int_0^{\infty} -\alpha t^{\alpha-1}e^{-t}dt\right) \\ &= (-0-(-0)) + \int_0^{\infty} \alpha t^{\alpha-1}e^{-t}dt = \alpha\int_0^{\infty} t^{\alpha-1}e^{-t}dt = \alpha\Gamma(\alpha) \end{aligned}$$

And so we proved the first part. We will prove inductively the following claim for the second part

Claim 2. For every $n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$.

Proof. For the base case $n = 1$, we have that

$$\Gamma(1) = \int_0^{\infty} t^0 e^{-t}dt = \left[-e^{-t}\right]_0^{\infty} = -(0-1) = 1 = 0!$$

For the base case assume truth for n, and notice that $\Gamma(n+1) = n\Gamma(n) = n \cdot (n-1)! = n!$, this is exactly what we want to prove and thus this concludes the claim. ■

■

3. We will use the following substitution of variables

$$t = \frac{x}{\beta} \text{ with } dt = \frac{1}{\beta}dx$$

Importantly we have that $t > 0 \Leftrightarrow x > 0$, and so the support of the distribution is still the same. Additionally

$$\begin{aligned}\frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)}dt \cdot \mathbb{1}_{\{t>0\}} &= \frac{((\frac{x}{\beta})^{\alpha-1}e^{-x/\beta})}{\Gamma(\alpha)}\frac{1}{\beta}dx \cdot \mathbb{1}_{\{x>0\}} \\ &= \frac{1}{\Gamma(\alpha)\beta^{\alpha-1}} \cdot \frac{1}{\beta} \cdot x^{\alpha-1}e^{-x/\beta}dx \cdot \mathbb{1}_{\{x>0\}} \\ &= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \cdot x^{\alpha-1}e^{-x/\beta}dx \cdot \mathbb{1}_{\{x>0\}}\end{aligned}$$

This means that

$$\int_0^\infty \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \cdot x^{\alpha-1}e^{-x/\beta}dx \cdot \mathbb{1}_{\{x>0\}} = \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \cdot x^{\alpha-1}e^{-x/\beta}dx = \int_0^\infty \frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)}dt = 1$$

Additionally it is obvious that the function is non-negative, and so we have that that gives us what we are looking for -

$\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{-x/\beta} \cdot \mathbb{1}_{\{x>0\}}$ is a distribution. (In particular, the Gamma distribution.)

We will denote

$$g(x) := \frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{-x/\beta} \cdot \mathbb{1}_{\{x>0\}}$$

For the rest of the exercise. ■

4. Notice that

$$\begin{aligned}\frac{1}{\beta} \cdot \mathbb{E}[X] &= \frac{1}{\beta} \cdot \int_0^\infty x \cdot g_X(x)dx \\ \text{(Variable Change } &= [x = \beta t,\ dx = \beta dt]) \\ &= \frac{1}{\beta}\int_0^\infty \beta t f(t)dt = \frac{\alpha\Gamma(\alpha)}{\Gamma(\alpha)} = \alpha\end{aligned}$$

And so

$$\mathbb{E}[X] = \alpha\beta$$

Similarly we can calculate

$$\begin{aligned}\frac{1}{\beta^2}\mathbb{E}[X^2] &= \frac{1}{\beta^2} \cdot \int_0^\infty x^2 \cdot g_X(x)dx = \frac{1}{\beta^2}\int_0^\infty \beta^2 t^2 f(t)dt \\ &= \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} = \frac{1}{\Gamma(\alpha)} \cdot ((\alpha+1)\Gamma(\alpha+1)) = \frac{1}{\Gamma(\alpha)} \cdot (\alpha+1) \cdot \alpha\Gamma(\alpha) = \alpha(\alpha+1)\end{aligned}$$

Where we use exactly the same concepts as when we did for the calculation of the expected value. And so

$$\mathbb{E}[X^2] = \alpha(\alpha+1)\beta^2$$

And so

$$\boxed{\mathrm{Var}(X)} = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \alpha(\alpha+1)\beta^2 - (\alpha\beta)^2 = \boxed{\alpha\beta^2}$$

■

Take $c > 0$. Similar to before we get that

$$\left(\frac{1}{\beta}\right)^c \mathbb{E}[X^c] = \left(\frac{1}{\beta}\right)^c \int_0^\infty x^c g_X(x)dx$$
$$= \beta^{-c}\int_0^\infty \beta^c \cdot t^c f(t)dt = \int_0^\infty t^c f(t)dt = \frac{\Gamma(\alpha+c)}{\Gamma(\alpha)}$$

And so

$$\boxed{\mathbb{E}[X^c] = \beta^c \frac{\Gamma(\alpha+c)}{\Gamma(\alpha)}}$$

Just to be clear, throughout this whole exercise we are utilizing the change of variables we saw in the previous part - this, as can be seen in this exercise, really simplifies the problem. Notice that throughout this exercise we also use the following idea -

$$\int_0^\infty t^\gamma f(t)dt = \int_0^\infty t^\gamma \frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)}dt = \frac{1}{\Gamma(\alpha)}\int_0^\infty t^{\alpha+\gamma-1}e^{-t}dt = \frac{\Gamma(\alpha+\gamma)}{\Gamma(\alpha)}$$

5. Notice that we have 2 independent parameters α, β and so we have to check whether or not this is in the 2 dimensional family of exponential distribution. Remember the definition of the exponential family of distributions from the lecture -

Definition 1. Assume that Θ is an open subset of $\mathbb{R}^p$. We will say that a distribution $f_{\overrightarrow{\theta}}(\overrightarrow{y})$ is in the family of exponential distributions if all of the following occur:
(1). The support is independent of $\overrightarrow{\theta}$.
(2). The distribution can be written in the following way

$$f_{\overrightarrow{\theta}}(\overrightarrow{y}) = \exp\left(\sum_{j=1}^k c_j(\overrightarrow{\theta})T_j(\overrightarrow{y}) + d(\overrightarrow{\theta}) + S(\overrightarrow{y})\right)$$

Notice first that the support of the density function is independent of the parameters - (all of the non-negative numbers). And so what is left for us is to show that the density function of the Γ distribution can be expressed in the following way

$$f_{(\alpha,\beta)}(x) := \exp\left(\sum_{j=1}^k c_j(\alpha,\beta)T_j(x) + d(\alpha,\beta) + S(x)\right)$$

Now we can try to show this. Take $X \sim \Gamma(\alpha,\beta)$.

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta} = e^{-x/\beta}\cdot e^{-\log\Gamma(\alpha)-\alpha\log\beta+(\alpha-1)\log x}$$
$$= e^{-x/\beta+(\alpha-1)\log x-\log\Gamma(\alpha)-\alpha\log\beta}$$

Now take

$$\overrightarrow{c}(\alpha,\beta)=\left(\alpha-1,-\frac{1}{\beta}\right),\overrightarrow{T}(x)=(\log x,x),S(x):=0,d(\alpha,\beta)=-\log\Gamma(\alpha)-\alpha\log\beta$$

And we get that we do have the required structure which tells us that we have that the Γ distribution is in the 2-dimensional family of exponential distribution. ■

6. Remember the Fischer-Neyman Factorization theorem

> **Theorem 1** (Fisher-Neyman Factorization Theorem). *A statistic $T\left(\overrightarrow{Y}\right)$ is sufficient for a parameter θ if and only if for all $\theta\in\Theta$ we have: $L\left(\theta;\overrightarrow{y}\right)=g\left(T\left(\overrightarrow{y}\right)\mid\theta\right)\cdot h\left(\overrightarrow{y}\right)$. Where the first factor isn't directly dependent on $\overrightarrow{y}$ and the second factor is independent of θ.*

This means that we need to find the likelihood function and then try to factor it as in the theorem. We have i.i.d. samples and so

$$L(\alpha,\beta;\overrightarrow{x})=\prod_{i=1}^{n}f_{X_i}(x_i)=\prod_{i=1}^{n}\left(\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x_i^{\alpha-1}e^{-x_i/\beta}\cdot\mathbb{1}_{\{x_i>0\}}\right)=$$

$$=\left(\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\right)^{n}\cdot\mathbb{1}_{\{\forall i:x_i>0\}}\cdot\left(\prod_{i=1}^{n}x_i\right)^{\alpha-1}\cdot\prod_{i=1}^{n}e^{-x_i/\beta}=$$

$$=\left(\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\right)^{n}\cdot\mathbb{1}_{\{\forall i:x_i>0\}}\cdot\left(\prod_{i=1}^{n}x_i\right)^{\alpha-1}\cdot e^{-\frac{1}{\beta}\sum_{i=1}^{n}x_i}$$

And so it is natural to consider the statistic

$$T\left(\overrightarrow{X}\right):=\left(T_1\left(\overrightarrow{X}\right):=\sum_{i=1}^{n}X_i,T_2\left(\overrightarrow{X}\right):=\prod_{i=1}^{n}X_i\right)$$

Now we can see that

$$L(\alpha,\beta;\overrightarrow{x})=\left(\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\right)^{n}\cdot\mathbb{1}_{\{\forall i:x_i>0\}}\cdot\left(T_2\left(\overrightarrow{x}\right)\right)^{\alpha-1}\cdot e^{-\frac{1}{\beta}T_1\left(\overrightarrow{x}\right)}$$

And so the natural factorization is

$$g(T(\overrightarrow{X})\mid\overrightarrow{\theta}):=\left(\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\right)^{n}\cdot\left(T_2\left(\overrightarrow{x}\right)\right)^{\alpha-1}\cdot e^{-\frac{1}{\beta}T_1\left(\overrightarrow{x}\right)}$$
$$h(\overrightarrow{x}):=\mathbb{1}_{\{\forall i:x_i>0\}}$$

And so we have the factorization needed - T is a sufficient statistic. ■

7. Remember the factorization from the previous part, notice that when α is known that we can redo the factorization in the following manner

$$g(T(\overrightarrow{X})\mid\beta):=\left(\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\right)^{n}\cdot e^{-\frac{1}{\beta}T(\overrightarrow{X})}$$
$$h(\overrightarrow{x}):=\left(\prod_{i=1}^{n}x_i\right)^{\alpha-1}\cdot\mathbb{1}_{\{\forall i:x_i>0\}}$$

Where $T(\overrightarrow{X}):=\sum_{i=1}^{n}x_i$ is the statistic we are interested in. And so T is a sufficient statistic. Remember the following criterion for minimality

Assume T is a sufficient statistic. T is minimal sufficient if and only if:

The ratio of the likelihoods of 2 samples is independent of $\theta \iff$ T agrees on the samples

And so, by taking two samplings $(x_i)_{i=1}^n, (y_i)_{i=1}^n$ we have that the ratio of likelihoods is

$$\frac{L(\beta;\overrightarrow{x})}{L(\beta;\overrightarrow{y})} = \frac{\left(\frac{1}{\Gamma(\alpha)\beta^\alpha}\right)^n \cdot \mathbb{1}_{\{\forall i: x_i>0\}} \cdot \left(\prod_{i=1}^n x_i\right)^{\alpha-1} \cdot e^{-\frac{1}{\beta}\sum_{i=1}^n x_i}}{\left(\frac{1}{\Gamma(\alpha)\beta^\alpha}\right)^n \cdot \mathbb{1}_{\{\forall i: y_i>0\}} \cdot \left(\prod_{i=1}^n y_i\right)^{\alpha-1} \cdot e^{-\frac{1}{\beta}\sum_{i=1}^n y_i}} =$$

$$= \frac{\mathbb{1}_{\{\forall i: x_i>0\}}}{\mathbb{1}_{\{\forall i: y_i>0\}}} \cdot \left(\prod_{i=1}^n \left(\frac{x_i}{y_i}\right)\right)^{\alpha-1} \cdot e^{-\frac{1}{\beta}(\sum_i y_i - \sum_i x_i)}$$

Notice that the 2 leftmost factors in the above product are independent of β, and so the total ratio is independent of β if and only if the rightmost factor in the product is independent of β. The only way for the rightmost factor to be independent of β is for the exponent to be zero. This only happens when

$$\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

This tells us that the ratio is independent of β if and only if $T(\overrightarrow{X}) = T(\overrightarrow{Y})$ and so we have shown what we need in order to know whether T is a minimal sufficient statistic or not - in this case T is a minimal sufficient statistic. ■

8. First notice that

$$\underline{\Gamma\left(\frac{1}{2}\right)} = \int_0^\infty \frac{e^{-t}}{\sqrt{t}}dt$$

$$\text{(Variable Change } = [t = u^2/2,\ dt = udu])$$

$$= \sqrt{2}\cdot\int_0^\infty \frac{e^{-u^2/2}}{u}\cdot udu = \frac{2}{\sqrt{2}}\cdot\frac{\sqrt{\pi}}{\sqrt{\pi}}\cdot\int_0^\infty e^{-u^2/2}du$$

$$= \sqrt{\pi}\cdot\int_{-\infty}^\infty \frac{e^{-u^2/2}}{\sqrt{2\pi}}du = \sqrt{\pi}\cdot\int_{-\infty}^\infty f_{\mathcal{N}(0,1)}(u)du = \sqrt{\pi}\cdot 1 = \underline{\sqrt{\pi}}$$

Where F is the CDF of $\mathcal{N}(0,1)$. Now take $X \sim \mathcal{N}(0,1)$, and $Y = X^2$. We want to find the distribution of Y. Notice that if $Z \sim \Gamma(\frac{1}{2}, 2)$ then the probability density function is

$$f_Z(x) := \frac{1}{\Gamma(\frac{1}{2})2^{\frac{1}{2}}}\cdot x^{-\frac{1}{2}}e^{-\frac{1}{2}x}\cdot\mathbb{1}_{\{x>0\}} = \frac{1}{\sqrt{2\pi}}\cdot x^{-\frac{1}{2}}e^{-\frac{1}{2}x}\cdot\mathbb{1}_{\{x>0\}}$$

And so we want to show that $f_Z(x)dx = f_Y(x)dx$. This would imply $Y \sim \Gamma\left(\frac{1}{2}, 2\right)$. And so

$$\underline{f_Y(x)dx} = f_{X^2}(x)dx = f_{X^2}(\sqrt{x})\cdot\frac{1}{2\sqrt{x}}\cdot 2\cdot\mathbb{1}_{\{x>0\}}dx$$

$$= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x}\cdot\frac{1}{\sqrt{x}}\cdot\mathbb{1}_{\{x>0\}}dx = \underline{f_Z(x)dx} \implies Y \sim \Gamma\left(\frac{1}{2}, 2\right)$$

■

2 Second Problem

In a specific population, p of the people make more then 50,000$ a year. We are interested in estimating p. We go through the following process:

- Put 100 envelopes in a box in the following way - $100x$ of them contain the question "Is your annual salary higher than 50,000$ a year?" and the rest of the envelopes contain the question "Is your annual salary lower or equal to 50,000$ a year?"
- Choose N people in random from the population.
- Each person takes an envelope, reads the question returns the envelope and answers the question.
- Assume only the person who takes out the envelope knows which question they got, and that they answer honestly.

Denote by Y the number of people who answered yes, and p' the probability that the first person answers yes. Assume the x is known.

1. How does Y distribute?
2. Find an estimator for p using the moment method - $\hat{p}_{MME}$.
3. Calculate the mean and variance of $\hat{p}_{MME}$. Express what you think the value of x should be based on this.
4. Express the possible set of values of p' as a function of x.
5. Find a maximum likelihood estimator for p'.
6. Set $x = \frac{1}{3}$ (for this part only). Show that if only one person agreed to participate ($N = 1$) it is more optimal to use the constant estimator $\frac{1}{2}$ instead of the MLE found in the last part
7. Find an MLE estimator for p. Discuss the advantages and disadvantages of $\hat{p}_{\text{MLE}}$ versus $\hat{p}_{MME}$.

(This process is used to deal with sensitive information).

Solution:

1. Define by X_i the Bernoulli random variable of the answer of the i-th person being "yes". First, examine the case for the first person:

   Our goal is to find $\mathbb{P}(X_1 = 1)$. Now, define the following events:

   - $A := \{$the question in the envelope was: "Is your annual salary higher than 50,000$"$\}$.
   - $B := \{$the person which is picked has an annual salary higher than 50,000$\}$.

   Since we assume that the person is honest, we get that (by the law of total probability)

   $$\mathbb{P}(X = 1) = \mathbb{P}(A \cap B) + \mathbb{P}(\bar{A} \cap \bar{B}) \overset{(*)}{=} \mathbb{P}(A) \cdot \mathbb{P}(B) + \mathbb{P}(\bar{A}) \cdot \mathbb{P}(\bar{B})$$

   (** the people and the envelopes are chosen independently, therefore A and B are independent.*)
   We have that $\mathbb{P}(A) = \frac{100x}{100} = x$. Assuming that $p \in [0, 1]$, we have that $\mathbb{P}(B) = p$. Thus, we get that

   $$\mathbb{P}(X = 1) = p' = x \cdot p + (1 - x) \cdot (1 - p) = 2px + 1 - p - x = (2x - 1) \cdot p + 1 - x$$

Therefore $X \sim \text{Ber}(p')$.

Since each person returns the envelope he had chosen and the people are chosen independently, we get that for all $1 \leq i \leq N$, $X_i \sim Ber(p')$ i.i.d (where $p' = (2x-1) \cdot p + 1 - x$). Since Y is the sum of N i.i.d Bernoulli random variables, we conclude that $Y \sim \text{Bin}(N, p)$.

2. We will use the MME to get an estimator for p', denoted by $\hat{p'}_{MME}$, and derive the requested estimator from it ($\hat{p}_{MME}$). The Moment Method Estimation states that to estimate $p' = \mathbb{E}[Ber(p')]$, it suffices to estimate the first moment of $Ber(p')$, in our case $\hat{p'}_{MME} = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{Y}{N}$. Since $p' = (2x-1) \cdot p + 1 - x$, it holds that

$$\hat{p'}_{MME} = (2x-1) \cdot \hat{p}_{MME} + 1 - x$$

$$\overset{(**)}{\Longrightarrow} \hat{p}_{MME} = \frac{\hat{p'}_{MME} + x - 1}{2x-1} = \frac{Y/N + x - 1}{2x-1}$$

*** This is assuming that $x \neq 1/2$. If indeed $x = 1/2$, then we have that $p' = p + 1 - p - 1/2 = 1/2$ and thus does not depends on p and we could not estimate p using the Moment Method Estimation.*

3. Since $Y \sim Bin(p', N)$ we have that $\mathbb{E}[Y] = Np'$, $Var(Y) = Np'(1-p')$. Therefore:

   - $\mathbb{E}[\hat{p}_{MME}] = \mathbb{E}\left[\frac{\frac{Y}{N}+x-1}{2x-1}\right] \overset{\text{linearity of expc.}}{=} \frac{\frac{\mathbb{E}[Y]}{N}+x-1}{2x-1} = \frac{\frac{N \cdot p'}{N}+x-1}{2x-1} = \frac{2px+1-p-x+x-1}{2x-1} = p$
   - $Var(\hat{p}_{MME}) = Var\left(\frac{\frac{Y}{N}+x-1}{2x-1}\right) \overset{(***)}{=} Var\left(\frac{Y/N}{2x-1}\right) = \frac{Var(Y)}{N^2(2x-1)^2} = \frac{Np'(1-p')}{N^2(2x-1)^2} = \frac{p'(1-p')}{N(2x-1)^2} = \frac{((2x-1)\cdot p+1-x)\cdot(x-(2x-1)\cdot p)}{N(2x-1)^2} = \frac{(2x-1)^2 \cdot p(1-p)+(1-x)x}{N(2x-1)^2} = \frac{p(1-p)}{N} + \frac{(1-x)x}{N(2x-1)^2}$

   **** The variance is invariant to changes in the location parameter.*

   First, as stated in the last question (2), $x \neq \frac{1}{2}$.

   Recall the definition of the Mean Squared Error (MSE):

   Definition 2. Given an estimator $\hat{\theta}$ we define the Mean-Square-Error

   $$MSE(\hat{\theta}) = \mathbb{E}_\theta\left[(\hat{\theta} - \theta)^2\right]$$

   It can be seen (similar to how you calculate variance of a random variable) that this is equal to $\text{bias}^2(\hat{\theta}) + \text{Var}(\hat{\theta})$ Where the bias is $\text{bias}(\hat{\theta}) := \theta - \mathbb{E}[\hat{\theta}]$

   Where $\text{Bias}_p(\hat{p}_{MME}) = \mathbb{E}_p[\hat{p}_{MME}] - p$. Now, notice that $\hat{p}_{MME}$ is an unbiased estimator and therefore $\text{MSE}_p[\hat{p}_{MME}] = Var_p(\hat{p}_{MME})$. Moreover, the left term in $Var(\hat{p}_{MME})$ is exactly $Var\left(\frac{Z}{N}\right)$ where $Z \sim Bin(p, N)$, and is therefore independent of x. So in order to minimize the MSE and optimize the estimator, we need to minimize the right term of $Var(\hat{p}_{MME})$. We have that $\frac{(1-x)x}{N(2x-1)^2} \geq 0$ since $x \in [0,1]$, and it is easy to see that it is equal to 0 if and only if $x \in \{0,1\}$. This implies that we minimize the error if we ask the same question to everyone (thus avoiding the uncertainty of determining which question was asked).

4. First, as we saw in question 2, if $x = \frac{1}{2}$ then we have that $p' = \frac{1}{2}$ and does not depend on p and x. So we will consider now the case where $x \neq \frac{1}{2}$:

   - $\boldsymbol{x < \frac{1}{2}}$: Since $p \in [0,1], x < \frac{1}{2}$ we have that:

   $$x = 2x - 1 + 1 - x \leq p' = (2x-1) \cdot p + 1 - x \leq 1 - x$$

- $\boldsymbol{x > \frac{1}{2}}$: Since $p \in [0,1], x > \frac{1}{2}$ we have that:

$$1 - x \leq p' = (2x-1) \cdot p + 1 - x \leq 2x - 1 + 1 - x = x$$

Summarized, we have that:

$$p' \in \begin{cases} [x, 1-x] & 0 \leq x < 1/2 \\ [1-x, x] & 1/2 < x \leq 1 \\ \{1/2\} & x = 1/2 \end{cases}$$

5. First, notice that the likelihood of a single observation of a person's answer ($x_i \in \{0,1\}$) is $L(x_i|p') = (p')^{x_i} \cdot (1-p')^{1-x_i}$ (recall that the answers of the people are identical and independently distributed according to $Ber(p')$). Thus, the likelihood of N observations $\vec{x} = (x_1, ..., x_N)$ is the following:

$$L(p' \mid x_1, ..., x_N) \overset{X_i \text{ are i.i.d}}{=} \prod_{i=1}^{N} L(x_i|p') = \prod_{i=1}^{N} (p')^{x_i} \cdot (1-p')^{1-x_i} = (p')^{\sum_{i=1}^{N} x_i} \cdot (1-p')^{\sum_{i=1}^{N}(1-x_i)}$$
$$\overset{Y=\sum_{i=1}^{N} x_i}{=} (p')^{Y} \cdot (1-p')^{N-Y}$$

We want to estimate p' by $\hat{p'}_{MLE} = \text{argmax}_{p' \in [0,1]} L(p'|\vec{x})$. For simplicity, we will find the argmax of the log-likelihood function (log is a monotone-increasing function, therefore it has the same argmax as the original function).

$$l(\vec{x}|p') = \log\left((p')^{Y} \cdot (1-p')^{N-Y}\right) = Y \log(p') + (N-Y)\log(1-p')$$
$$\frac{\partial l}{\partial p'} = \frac{Y}{p'} - \frac{N-Y}{1-p'} = \frac{Y - p' \cdot Y + p' \cdot Y - p' \cdot N}{p'(1-p')} = \frac{Y - p' \cdot N}{p'(1-p')}$$

And it is equals to 0 if and only if $p' = \frac{Y}{N}$. It is easy to see that for $0 < p' < \frac{Y}{N}$ the derivative is positive and for $\frac{Y}{N} < p' < 1$ it is negative and thus it is the argmax of the log-likelihood. From the last question (4) we have that the value of p' depends on the value of x (the percentage of the envelopes with the question "Is your annual salary higher than 50,000$ a year?"), therefore we actually have that:

$$\hat{p'}_{MLE} = \begin{cases} x & (0 \leq x < 1/2 \wedge Y/N < x) \vee (1/2 < x \leq 1 \wedge Y/N > x) \\ Y/N & (0 \leq x < 1/2 \wedge x < Y/N < 1-x) \vee (1/2 < x \leq 1 \wedge 1-x < Y/N < x) \\ 1-x & (0 \leq x < 1/2 \wedge Y/N > 1-x) \vee (1/2 < x \leq 1 \wedge Y/N < 1-x) \\ 1/2 & x = 1/2 \end{cases}$$

6. Set $x = 1/3$. From question 4 we have that $1/3 \leq p' \leq 2/3$. If we sample only one individual (equivalent to only one person agreeing to participate), and thus $Y/N = X_1 \in \{0,1\}$ and therefore $Y/N < 1/3$ or $Y/N > 2/3$. From last question we get:

$$\hat{p'}_{MLE} = \begin{cases} 2/3 & \text{w.p. } \mathbb{P}(Y/N = X_1 = 1) = p' \\ 1/3 & \text{w.p. } \mathbb{P}(Y/N = X_1 = 0) = 1 - p' \end{cases}$$

(depends on the observation).
We compute and compare the MSE for each estimator.

- $\text{MSE}_{p'}[\hat{p'}_{MLE}]$:

$$\text{Bias}_{p'}(\hat{p'}_{MLE}) = \mathbb{E}\left[\hat{p'}_{MLE}\right] - p' = p' \cdot \frac{2}{3} + (1-p') \cdot \frac{1}{3} - p' = \frac{1}{3} - \frac{2}{3}p'$$

$$Var\left(\hat{p'}_{MLE}\right) = \mathbb{E}\left[\hat{p'}^2_{MLE}\right] - \left(\frac{1}{3} + \frac{1}{3}p'\right)^2 = \frac{1}{9} + \frac{1}{3}p' - \frac{1}{9} - \frac{2}{9}p' - \frac{1}{9}(p')^2 = -\frac{1}{9}(p')^2 + \frac{1}{9}p'$$

$$\implies \text{MSE}_{p'}[\hat{p'}_{MLE}] = Var(\hat{p'}_{MLE}) + \text{Bias}^2_{p'}(\hat{p'}_{MLE}) = \frac{1}{9} - \frac{1}{3}p' + \frac{1}{3}(p')^2$$

- $\text{MSE}_{p'}\left[\frac{1}{2}\right]$:

$$\text{Bias}_{p'}\left(\frac{1}{2}\right) = \frac{1}{2} - p', \quad Var_{p'}\left(\frac{1}{2}\right) = 0$$

$$\implies \text{MSE}_{p'}\left[\frac{1}{2}\right] = \left(\frac{1}{2} - p'\right)^2 = \frac{1}{4} - p + p^2$$

It holds that:

$$\begin{aligned}\text{MSE}_{p'}\left[\frac{1}{2}\right] < \text{MSE}_{p'}\left[\hat{p'}_{MLE}\right] &\iff \frac{1}{4} - p' + (p')^2 < \frac{1}{9} - \frac{1}{3}p' + \frac{1}{3}(p')^2 \\ &\iff 0 < -\frac{5}{36} + \frac{2}{3}p' - \frac{2}{3}(p')^2 \\ &\iff 0.295 < p' < 0.704\end{aligned}$$

And since $1/3 \le p' \le 2/3$ we proved the requested. ■

7. Recall that $p' = (2x-1)\cdot p + 1 - x$. This yields $\hat{p}_{MLE} = \frac{x-1+\hat{p'}_{MLE}}{2x-1}$. From question 5 we get:

$$\hat{p}_{MLE} = \begin{cases} 1 & (0 \le x < 1/2 \wedge Y/N < x) \vee (1/2 < x \le 1 \wedge Y/N > x) \\ \hat{p}_{MME} = \frac{x-1+Y/N}{2x-1} & (0 \le x < 1/2 \wedge x < Y/N < 1-x) \vee (1/2 < x \le 1 \wedge 1-x < Y/N < x) \\ 0 & (0 \le x < 1/2 \wedge Y/N > 1-x) \vee (1/2 < x \le 1 \wedge Y/N < 1-x) \end{cases}$$

In question 3 we saw that $\hat{p}_{MME}$ is unbiased, and since $\hat{p}_{MLE}$ is not always equal to it, then it is not always unbiased. Moreover, notice that whenever Y/N is outside of the boundaries then $\hat{p}_{MME}$ can be either < 0 or > 1, which causes a bigger error whereas $\hat{p}_{MLE}$ is either 0 or 1. Therefore, $\hat{p}_{MLE}$ induces less error and thus preferable to use.

---

3 Third Problem

Assume (X, Y) distributes with joint density function $f(x, y \mid \theta) = e^{-(x\theta + \frac{1}{\theta}y)}$. (Slight correction: the function as it appears here is not strictly a distribution, that is because the integral over all of the plane is infinite. after talking to the T.A. it was corrected to $f'(x,y) := e^{-(x\theta+\frac{1}{\theta}y)} \cdot \mathbb{1}_{\{x,y>0\}}$)

1. Show that the Fischer information on n samples is $I(\theta) = \frac{2n}{\theta^2}$
2. Show that (T, U) is a sufficient statistic, yet is not complete where

$$T := \sqrt{\sum_i Y_i / \sum_i X_i}, \text{ and } U := \sqrt{\sum_i Y_i \sum_i X_i}$$

3. Show that T is an MLE for θ
4. Is T a UMVUE? Can we know whether there is a UMVUE using Lehman-Scheffe?
5. Calculate the first 2 moments $\mathbb{E}[T], \mathbb{E}[T^2]$ (Hint: Use the gamma function)
6. Define
$$z_1 := (n-1)/\sum_i X_i, \text{ and } z_2 := \sum_i Y_i/n$$
Show that they are unbiased estimators for θ, and calculate their variance.
7. Find the best estimator of the form $\alpha z_1 + (1-\alpha)z_2$ What is it's variance? Compare to T's variance after the removal of the bias.

Solution:

The first thing we will do is notice that we can decompose the distribution in the following way

$$f(x,y) = \left(\theta \cdot e^{-x\theta} \cdot \mathbb{1}_{\{x>0\}}\right) \cdot \left(\frac{1}{\theta} \cdot e^{-y/\theta} \cdot \mathbb{1}_{\{y>0\}}\right)$$

And so notice that if we take $Z_1 \sim \text{Exp}(\theta), Z_2 \sim \text{Exp}(\frac{1}{\theta})$ independently distributed random variables, we have that the joint density function of (Z_1, Z_2) is exactly that of our original (X, Y). This means that

$$(X,Y) \overset{d}{\equiv} (Z_1, Z_2)$$

That means we can think of X as an $\text{Exp}(\theta)$ random variable and that it is independent of Y which is an $\text{Exp}(\frac{1}{\theta})$ random variable. We will use this throughout the solution. Now we will return to the solution of the problem.

1. Remember that under enough regularity assumptions we have that the Fischer-information is
$$I(\theta) = -\mathbb{E}\left[\frac{d^2}{d\theta^2}\log\left(f_\theta(\overrightarrow{y})\right)\right]$$
(The regularity assumptions are basic ones that a simple exponential function obviously satisfies, like the twice differentiability criteria or being able to change order between differentiation and integration and the support being independent of θ.) First we will find $f_\theta(\overrightarrow{y})$.
$$\begin{aligned}
f_\theta(\overrightarrow{y}) &= \prod_{i=1}^n e^{-(x_i\theta + y_i/\theta)} = e^{-\theta\sum_{i=1}^n x_i - \frac{1}{\theta}\sum_{i=1}^n y_i} \\
\log f_\theta(\overrightarrow{y}) &= -\theta\sum_{i=1}^n x_i - \frac{1}{\theta}\sum_{i=1}^n y_i \\
\frac{d}{d\theta}\log f_\theta(\overrightarrow{y}) &= -\sum_{i=1}^n x_i + \frac{1}{\theta^2}\sum_{i=1}^n y_i \\
\frac{d^2}{d\theta^2}\log f_\theta(\overrightarrow{y}) &= -2\theta^{-3}\sum_{i=1}^n y_i
\end{aligned}$$
And so
$$I(\theta) = -\mathbb{E}\left[\frac{d^2}{d\theta^2}\log f_\theta(\overrightarrow{y})\right] = \mathbb{E}\left[2\theta^{-3}\sum_{i=1}^n y_i\right] = 2\theta^{-3}\mathbb{E}\left[\sum_{i=1}^n y_i\right] = 2\theta^{-3}\sum_{i=1}^n \mathbb{E}[y_i] = 2\theta^{-3}\cdot n\cdot\theta = \frac{2n}{\theta^2}$$

$$\left(\underline{\mathbb{E}[y_i]} = \int_0^\infty \frac{1}{\theta} \cdot te^{-\frac{1}{\theta}t}dt = \frac{1}{\theta}\left(\left[-t\theta e^{-\frac{1}{\theta}y}\right]_0^\infty + \int_0^\infty \theta e^{-\frac{1}{\theta}t}dt\right) = \frac{1}{\theta}\cdot 0 + \frac{1}{\theta}\left[-\theta^2 e^{-\frac{1}{\theta}t}\right]_0^\infty = \underline{\theta}\right)$$

■

2. First notice that $\frac{U}{T} = \sum_{i=1}^n X_i$, and $T \cdot U = \sum_{i=1}^n Y_i$. Now remember the Fisher-Neyman factorization theorem (see theorem 1), we will to try to factor the likelihood function -

$$L(\theta; (\overrightarrow{x}, \overrightarrow{y})) := \prod_{i=1}^n f_\theta(x_i, y_i) = \prod_{i=1}^n e^{-x_i\theta - \frac{1}{\theta}y_i} = e^{-\theta\sum_{i=1}^n x_i - \frac{1}{\theta}\sum_{i=1}^n y_i} = e^{-\theta\frac{U}{T} - \frac{1}{\theta}UT}$$

Now we want to use the statistic $M = (U, T)$ for factorization. This means taking

$$g\left(M\left(\overrightarrow{X}, \overrightarrow{Y}\right); \theta\right) := e^{-\theta\frac{U}{T} - \frac{1}{\theta}UT}$$
$$h\left(\overrightarrow{X}, \overrightarrow{Y}\right) := 1$$

Now using the factorization theorem we have that $M = (T, U)$ is a sufficient statistic. Now we also want to show that the statistic is not complete. Recall what a complete statistic is

> A statistic T is said to be complete if and only if for all measurable functions g:
> $$\forall\theta:\ \mathbb{E}\left[g(T(\overrightarrow{Y}))\right] = 0 \Longrightarrow \forall\theta:\ \mathbb{P}_\theta\left(g(T(\overrightarrow{Y})) = 0\right) = 1$$

This implies that if we find a measurable function g on M such that

$$\forall\theta:\ \mathbb{E}\left[g(M(\overrightarrow{X}, \overrightarrow{Y}))\right] = 0 \text{ and } \mathbb{P}_\theta\left(g(T(\overrightarrow{Y})) = 0\right) \neq 1$$

then we are done. Notice that

$$\mathbb{E}\left[U^2\right] = \mathbb{E}\left[\sum_i X_i \cdot \sum_i Y_i\right] = \sum_{i,j}\mathbb{E}[X_i \cdot Y_j] = n^2\mathbb{E}[XY] = n^2\mathbb{E}[X]\mathbb{E}[Y] = n^2\theta\frac{1}{\theta} = n^2$$

Where $\mathbb{E}[X] = \frac{1}{\theta}$, for the same reasons that $\mathbb{E}[Y] = \theta$. And so if we define $g(T, U) = U^2 - n^2$, then we have

$$\mathbb{E}\left[g\left(T, U\right)\right] = \mathbb{E}\left[U^2 - n^2\right] = \mathbb{E}\left[U^2\right] - n^2 = n^2 - n^2 = 0$$

Notice that U^2 is not constant and so g is not constant meaning we found a function g that satisfies what we are looking for: $M = (T, U)$ is not a complete statistic. ■

3. Remember from the calculation of the fischer information that the log-likelihood is

$$l\left(\theta; (\overrightarrow{x}, \overrightarrow{y})\right) = -\theta\sum_{i=1}^n x_i - \frac{1}{\theta}\sum_{i=1}^n y_i$$

Now we want to maximize this function (maximizing the likelihood is equivalent to maximizing the log-likelihood, since the log is monotone increasing), and so derive

$$\frac{d}{d\theta}l(\theta; (\overrightarrow{x}, \overrightarrow{y})) = -\sum_{i=1}^n x_i + \frac{1}{\theta^2}\sum_{i=1}^n y_i = 0 \Longrightarrow \theta^2 \cdot \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \Longrightarrow$$
$$\Longrightarrow \theta^2 = T^2 \Longrightarrow \underline{\theta = \pm T}$$

Of course T is positive and $\theta > 0$ and so the only option is $\theta = T$. Now we will check whether or not this is a maximum. The second derivative is

$$\frac{d^2}{d\theta^2} l(\theta; (\overrightarrow{x}, \overrightarrow{y})) = -\frac{2}{\theta^3} \sum_{i=1}^{n} y_i < 0$$

This is because $y_i > 0, \theta > 0$. And so if we take $\hat{\theta} = T$. We have a value of theta that maximizes the log-likelihood. And so T is an MLE estimator for θ. ■

4. Remember the following definition of a UMVUE -

Definition 3. An estimator $\hat{\theta}$ is called an unbiased estimator with uniformly minimal variance (UMVUE) if it satisfies the following conditions

(a) $\hat{\theta}$ is an unbiased estimator.

(b) For any other unbiased estimator $\hat{\theta_1}$

$$\text{Var}(\hat{\theta}) \leq \text{Var}(\hat{\theta_1})$$

For all $\theta \in \Theta$.

If we determine that T is a biased estimator for θ then we get that it is also not a UMVUE. This means showing that $\mathbb{E}[T] \neq \theta$. Notice that because $\{X_i\} \sim \text{Exp}(\theta)$ i.i.d. and $\{Y_i\} \sim \text{Exp}(\frac{1}{\theta})$ i.i.d. that we also have that

$$R = \sum_{i=1}^{n} X_i \sim \Gamma\left(n, \frac{1}{\theta}\right), \text{ and } S = \sum_{i=1}^{n} Y_i \sim \Gamma(n, \theta) \Longrightarrow T = \sqrt{\frac{S}{R}}$$

And so

$$\begin{aligned}
\mathbb{E}[T] = \mathbb{E}\left[\sqrt{\frac{S}{R}}\right] &= \int_0^\infty \int_0^\infty \sqrt{\frac{s}{r}} \cdot \frac{r^{n-1}\theta^n e^{-\theta r}}{\Gamma(n)} \cdot \frac{s^{n-1} e^{-\frac{1}{\theta}s}}{\Gamma(n)\theta^n} dr ds \\
&= \frac{\Gamma(n-\frac{1}{2})\Gamma(n+\frac{1}{2})}{\Gamma(n)^2} \cdot \theta^{\frac{1}{2}} \cdot \theta^{\frac{1}{2}} \left(\int_0^\infty \frac{r^{n-\frac{1}{2}-1}\theta^{n-\frac{1}{2}} e^{-\frac{1}{\theta}r}}{\Gamma(n-\frac{1}{2})} dr\right) \cdot \left(\int_0^\infty \frac{s^{n+\frac{1}{2}-1} e^{-\theta s}}{\Gamma(n+\frac{1}{2})\theta^{n+\frac{1}{2}}} ds\right) \\
&= \frac{\Gamma(n-\frac{1}{2})\Gamma(n+\frac{1}{2})}{\Gamma(n)^2} \cdot \theta \cdot \left(\int_0^\infty f_{\Gamma(n-\frac{1}{2}, \frac{1}{\theta})}(r) dr\right) \cdot \left(\int_0^\infty f_{\Gamma(n+\frac{1}{2}, \theta)}(s) ds\right) \\
&= \frac{\Gamma(n-\frac{1}{2})\Gamma(n+\frac{1}{2})}{\Gamma(n)^2} \cdot \theta
\end{aligned}$$

Notice that this means that $\mathbb{E}[T] \neq \theta$, and so T is biased and not a UMVUE. ■

Recall the Lehmann-Scheffe theorem -

Theorem 2 (Lehmann - Scheffe)**.** *If W is a sufficient and complete statistic for θ, and T is an unbiased estimator then*

$$T_1 = \mathbb{E}[T \mid W]$$

Is a UMVUE for θ.

Meaning to invoke this theorem we need to have the existence of a sufficient and complete statistic. Remember from the lecture that a complete and sufficient statistic is also minimal, and so if we find

a minimal sufficient statistic that isn't complete this means that there isn't a sufficient and complete statistic (These statistics should be functions of eachother which would mean that our non-complete minimal sufficient statistic should also be complete in contradiction).

This means that if we can show that M is minimal then since it is not complete there aren't sufficient and complete statistics which would tell us we cannot use the Lehmann-Scheffe theorem. Remember from previous problems the criteria for minimality of a sufficient statistic -

The ratio of the likelihoods of 2 samples is independent of $\theta \iff$ M agrees on the samples

Take 2 samplings $(\overrightarrow{x_1}, \overrightarrow{y_1}), (\overrightarrow{x_2}, \overrightarrow{y_2})$. Denote by (T_1, U_1), (T_2, U_2) the corresponding statistics. The ratio of the likelihoods is

$$\frac{L(\theta; (\overrightarrow{x_1}, \overrightarrow{y_1}))}{L(\theta; (\overrightarrow{x_2}, \overrightarrow{y_2}))} = \frac{e^{-\theta \frac{U_1}{T_1}} - \frac{1}{\theta} U_1 T_1}{e^{-\theta \frac{U_2}{T_2}} - \frac{1}{\theta} U_2 T_2} = e^{-\theta\left(\frac{U_1}{T_1} - \frac{U_2}{T_2}\right) - \frac{1}{\theta}(T_1 U_1 - T_2 U_2)}$$

For this to be independent of θ we need that

$$\left(\frac{U_1}{T_1} - \frac{U_2}{T_2}\right) = 0, \text{ and } T_1 U_1 - T_2 U_2 = 0$$

And so we have that $T_1 U_1 = T_2 U_2$, and $T_1 U_2 = T_2 U_1$. Notice that these are all positive and so we have that this is true if and only if $T_1 = T_2$, and $U_1 = U_2$. And so we have all together that

The ratio of the likelihoods of 2 samples is independent of $\theta \iff$ M agrees on the samples

Meaning we have proven that M is minimal sufficient yet not complete and for the reasons stated above we have that the theorem cannot be invoked in this case. ■

5. Remember that we already calculated $\mathbb{E}[T]$ -

$$\mathbb{E}[T] = \frac{\Gamma(n - \frac{1}{2})\Gamma(n + \frac{1}{2})}{\Gamma(n)^2} \cdot \theta$$

So now what is left is to calculate $\mathbb{E}\left[T^2\right]$, we will do this in a very similar way -

$$\begin{aligned}
\mathbb{E}\left[T^2\right] = \mathbb{E}\left[\frac{S}{R}\right] &= \int_0^\infty \int_0^\infty \frac{s}{r} \cdot \frac{r^{n-1}\theta^n e^{-\theta r}}{\Gamma(n)} \cdot \frac{s^{n-1} e^{-\frac{1}{\theta}s}}{\Gamma(n)\theta^n} ds dr \\
&= \frac{\Gamma(n-1)\Gamma(n+1)}{\Gamma(n)^2} \cdot \theta \cdot \theta \cdot \left(\int_0^\infty \frac{r^{n-2}\theta^{n-1} e^{-\theta r}}{\Gamma(n-1)} dr\right) \cdot \left(\int_0^\infty \frac{s^n e^{-\frac{1}{\theta}s}}{\Gamma(n+1)\theta^{n+1}} ds\right) \\
&= \frac{\Gamma(n-1) \cdot (n(n-1)\Gamma(n-1))}{(n-1)^2 \Gamma(n-1)^2} \cdot \theta^2 \cdot \left(\int_0^\infty f_{\Gamma(n-1, \frac{1}{\theta})}(r) dr\right) \cdot \left(\int_0^\infty f_{\Gamma(n+1, \theta)}(s) ds\right) \\
&= \frac{n}{n-1}\theta^2
\end{aligned}$$

6. We are supposed to show that $\mathbb{E}[z_1] = \mathbb{E}[z_2] = \theta$ -

$$\begin{aligned}
\mathbb{E}[z_1] = \mathbb{E}\left[\frac{n-1}{R}\right] &= \int_0^\infty \frac{n-1}{r}\frac{r^{n-1}\theta^n e^{-\theta r}}{\Gamma(n)}dr = \frac{(n-1)\Gamma(n-1)}{\Gamma(n)}\int_0^\infty \frac{r^{n-2}\theta^{n-1}e^{-\theta r}}{\Gamma(n-1)}dr \\
&= \frac{\Gamma(n)}{\Gamma(n)}\cdot\theta\cdot\left(\int_0^\infty f_{\Gamma(n-1,\frac{1}{\theta})}(r)dr\right) = \theta \\
\mathbb{E}[z_2] = \mathbb{E}\left[\frac{S}{n}\right] &= \int_0^\infty \frac{s}{n}\cdot\frac{s^{n-1}e^{-\frac{1}{\theta}s}}{\Gamma(n)\theta^n}ds = \frac{\Gamma(n+1)}{n\Gamma(n)}\theta\cdot\int_0^\infty \frac{s^n e^{-\frac{1}{\theta}s}}{\Gamma(n+1)\theta^{n+1}}ds \\
&= \frac{\Gamma(n+1)}{\Gamma(n+1)}\cdot\theta\cdot\left(\int_0^\infty f_{\Gamma(n+1,\theta)}(s)ds\right) = \theta
\end{aligned}$$

And so we have that they are unbiased estimators for θ. The second moments are -

$$\begin{aligned}
\mathbb{E}\left[z_1^2\right] = \mathbb{E}\left[\frac{(n-1)^2}{R^2}\right] &= \int_0^\infty \frac{(n-1)^2}{r^2}\frac{r^{n-1}\theta^n e^{-\theta r}}{\Gamma(n)}dr = (n-1)^2\frac{\Gamma(n-2)}{\Gamma(n)}\theta^2\left(\int_0^\infty \frac{r^{n-3}\theta^{n-2}e^{-\theta r}}{\Gamma(n-2)}dr\right) \\
&= \frac{(n-1)^2}{(n-2)(n-1)}\theta^2\cdot\left(\int_0^\infty f_{\Gamma(n-2,\frac{1}{\theta})}(r)dr\right) = \frac{n-1}{n-2}\theta^2 \\
\mathbb{E}\left[z_2^2\right] = \mathbb{E}\left[\frac{S^2}{n^2}\right] &= \int_0^\infty \frac{s^2}{n^2}\cdot\frac{s^{n-1}e^{-\frac{1}{\theta}s}}{\Gamma(n)\theta^n}ds = n^{-2}\frac{\Gamma(n+2)}{\Gamma(n)}\theta^2\cdot\left(\int_0^\infty \frac{s^{n+1}e^{-\frac{1}{\theta}s}}{\Gamma(n+2)\theta^{n+2}}ds\right) \\
&= \frac{n(n+1)}{n^2}\theta^2\cdot\left(\int_0^\infty f_{\Gamma(n+2,\theta)}(s)ds\right) = \frac{n+1}{n}\theta^2
\end{aligned}$$

And so the variances are

$$\begin{aligned}
\mathrm{Var}(z_1) &= \mathbb{E}\left[z_1^2\right] - \mathbb{E}[z_1]^2 = \frac{n-1}{n-2}\theta^2 - \theta^2 = \frac{1}{n-2}\theta^2 \\
\mathrm{Var}(z_2) &= \mathbb{E}\left[z_2^2\right] - \mathbb{E}[z_2]^2 = \frac{n+1}{n}\theta^2 - \theta^2 = \frac{1}{n}\theta^2
\end{aligned}$$

7. Notice that z_1 and z_2 distribute independently and this is because R and S distribute independently. And so if we denote $z = z_1 + z_2$. First notice that for every α this is an unbiased estimator for θ, thus our objective would be to just minimize $\mathrm{Var}(z)$.

$$\mathrm{Var}(z) = \mathrm{Var}(\alpha z_1) + \mathrm{Var}((1-\alpha)z_1) = \alpha^2\mathrm{Var}(z_1) + (1-\alpha)^2\mathrm{Var}(z_2) = \theta^2\left(\frac{\alpha^2}{n-2} + \frac{(1-\alpha)^2}{n}\right)$$

We want to minimize this value, and so we will derive it according to α -

$$\frac{d}{d\alpha}\mathrm{Var}(z) = \theta^2\left(\frac{2\alpha}{n-2} - \frac{2(1-\alpha)}{n}\right) = 0 \Longrightarrow \underline{\alpha = \frac{n-2}{2n-2}}$$

And so since $\frac{d}{d\alpha}\mathrm{Var}(z) = \theta^2\left(\frac{2}{n-2} + \frac{2}{n}\right) > 0$, this means the above α minimizes the variance. And so the best estimator of this form and it's variance is

$$z = \frac{1}{2}\left(\frac{n-2}{\sum_i X_i} + \frac{\sum_i Y_i}{n-1}\right) \Longrightarrow \mathrm{Var}(z) = \frac{(n-2)\theta^2}{(2n-2)^2} + \frac{n\theta^2}{(2n-2)^2} = \frac{\theta^2}{2n-2}$$

Remembering $\mathbb{E}[T]$, in order to remove the bias we need to define $\hat{T} := \frac{\Gamma(n)^2}{\Gamma(n-\frac{1}{2})\Gamma(n+\frac{1}{2})}T$ to remove the

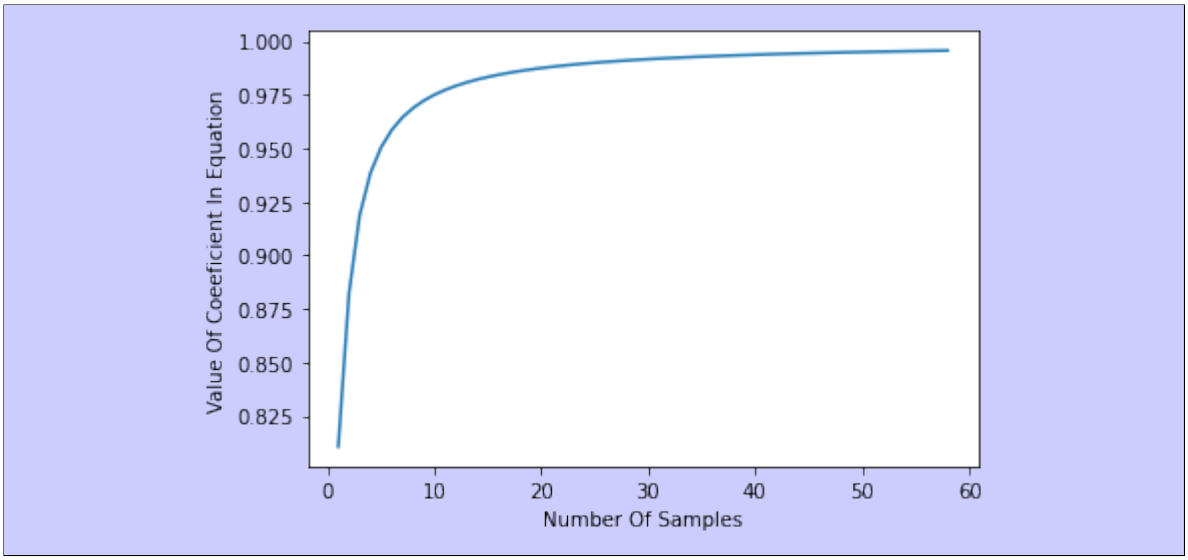
bias. Now this means we want to compare the variance of $\hat{T}$ to the variance of the above z -

$$\mathbb{E}\left[\hat{T}^2\right] = \frac{\Gamma(n)^4}{\Gamma(n-\frac{1}{2})^2\Gamma(n+\frac{1}{2})^2}\mathbb{E}\left[T^2\right] = \frac{n\Gamma(n)^4\theta^2}{(n-1)\Gamma(n-\frac{1}{2})^2\Gamma(n+\frac{1}{2})^2} \Longrightarrow$$

$$\Longrightarrow \mathrm{Var}\left(\hat{T}\right) = \mathbb{E}\left[\hat{T}^2\right] - \mathbb{E}\left[\hat{T}\right]^2 = \frac{\theta^2}{2(n-1)} \cdot 2\left(\frac{n\Gamma(n)^4}{\Gamma(n-\frac{1}{2})^2\Gamma(n+\frac{1}{2})^2} - (n-1)\right)$$

$$= \mathrm{Var}(z) \cdot 2\left(\frac{n\Gamma(n)^4}{\Gamma(n-\frac{1}{2})^2\Gamma(n+\frac{1}{2})^2} - (n-1)\right)$$

This is the plot of the coefficient function, this is done in the beginning of the code file submitted. There also is an explanation there for why this means T is better than z as an estimate for θ.



---

4 Fourth Problem

We say that a random variable X distributes Pareto, denoting $X \sim \mathrm{Par}(x_m, \alpha)$, if the distribution function of X is

$$f_X(x) := \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, & x \geq x_m \\ 0, & \text{Otherwise} \end{cases}$$

1. Sample $n = 1,000$ samples from a pareto distribution with $x_m = 5, \alpha = 7$. Show a histogram of these samplings.
2. Write a function that calculates the likelihood of a single sampling for $x_m = 5$ and an unknown α. Use this to calculate the general log likelihood of the whole sampling.
3. Numerically find the maximum log-likelihood.
4. Go over this process 10,000 times and find confidence intervals with significance $0.99, 0.95$. Show a diagram with 100 confidence intervals with significance 0.95, and compare to the real parameter value. Discuss the result.

5. Set $\alpha = 1.1$ (leave x_m as is). Estimate numerically the expected value using 100 sampling of sizes

$$N \in \{1, 1,00, 10, 000, 1, 000, 000, 100, 000, 000\} = \{10^0, 10^2, 10^4, 10^6, 10^8\}$$

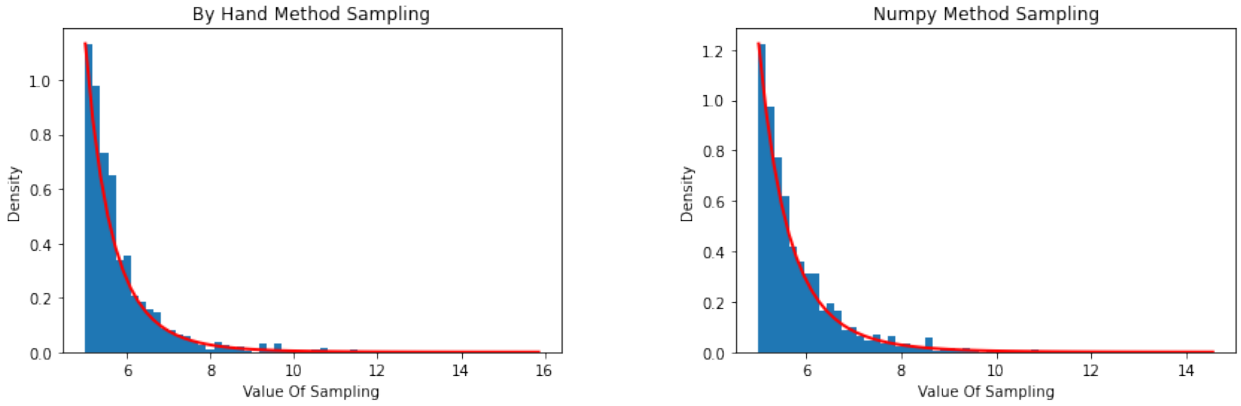
Present the results of these 500 samplings. Do the same with $\alpha = 0.9$. Discuss this result

---

Solution:

The code we are going to use was submitted in an ipynb (jupyter notebook) file. The file supports LaTeX but it isn't easily readable. Therefore we also submit a PDF or .html version of the notebook where it will be easier to read, especially the mathematical parts. Here we will summarize the main results/concepts that we want to present from the code file.

1. In the code file we presented 2 different methods for sampling the pareto distributions. The plots you get from the samplings look like this -



(a) Method Of Sampling "By Hand"

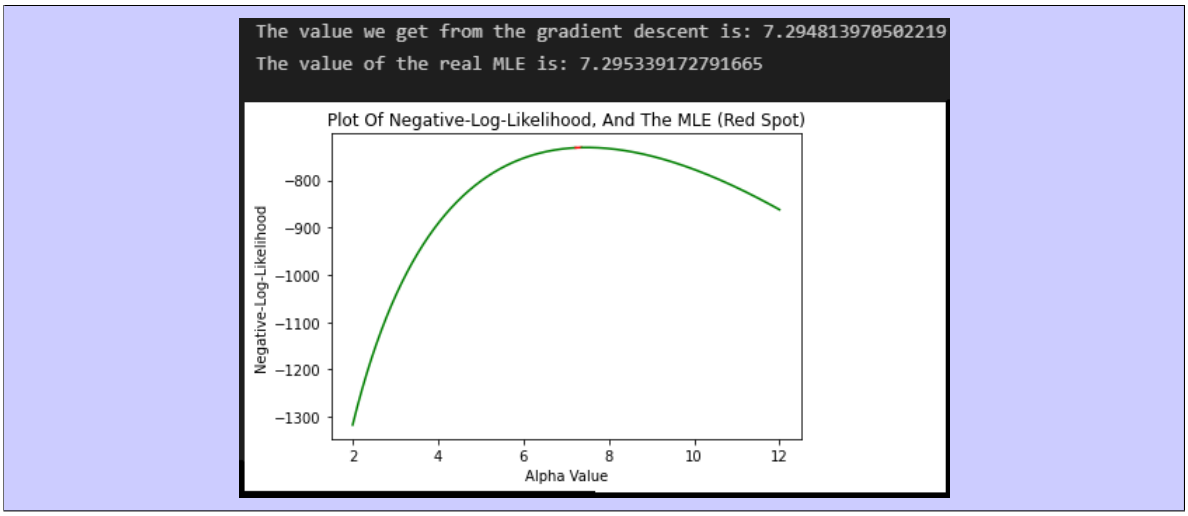
(b) Using The Built-In Numpy Function

Figure 1: Both Sampling Methods

Where the red line is the real density function of the distribution, And the blue rectangles are the bins of the histogram of the sampling we got. From these images alone it can be seen that these methods are equivalent (in the sense that both sample from the same distribution). ■

2. The solution is fully presented in the code file. Mostly straight-forward and simple code. As an example of this calculation the log-likelihood of the sampling we did using the numpy method. One of the times this came out to be -767.504510428529. Of course you can check the notebook that we submitted for the result from the last time it ran before we gave in the project. ■

3. We use 2 methods to find the maximum. One numerically (as was asked for), and one analytically. The numerical method uses the gradient descent method (which is described in the jupyter notebook) for the negative log-likelihood (which maximizes our original function), and the analytical solution uses deriving the function and comparing it to 0. We then compare these different solutions. When running the code it can be very quickly seen that these methods lead to *very* close values of α. We get the following plot of the log-likelihood and we can see how close our solution is to the real value -

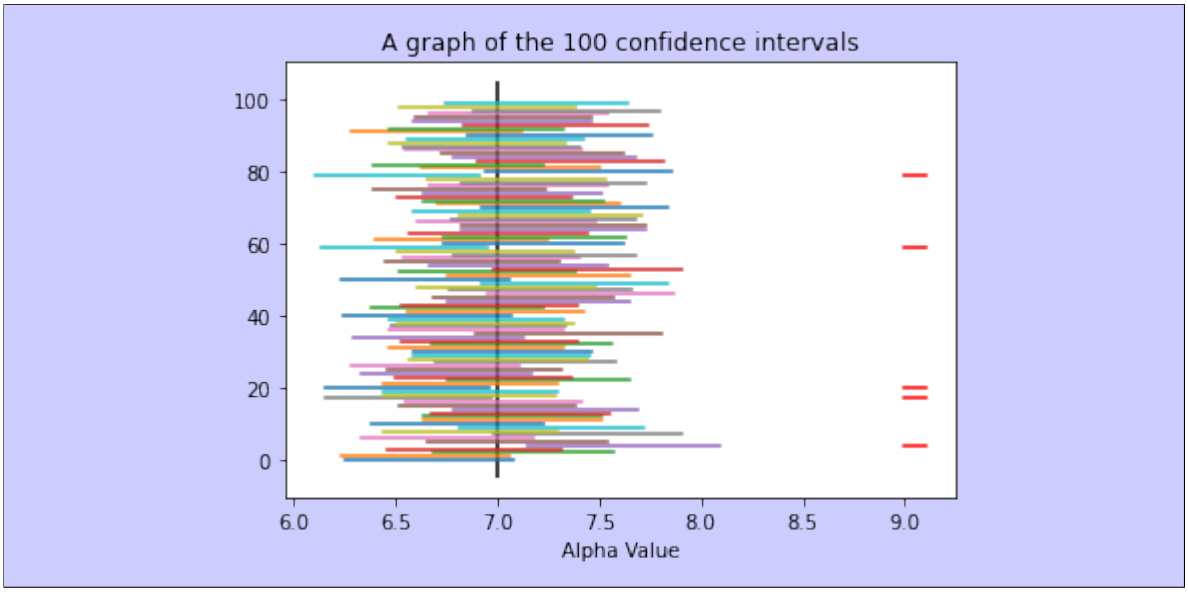


4. We presented 2 methods in the code. One is based on using MLE-s: if we sample enough MLE-s we can take a confidence interval for the MLE-s which will also serve as a confidence interval for the "real" α. The other is based on using pivots. For the pivot we use the following quantity

$$2\alpha \sum_{i=1}^{n} \log\left(\frac{x}{x_m}\right) \sim \chi^2_{2n}$$

In the code file there is an explanation for why this is true. We also present a comparison between these methods according to some basic metrics on the intervals - the average length of the intervals, the standard deviation and how much these intervals are symmetric according to the real α.

We then presented 100 confidence intervals according to the pivot method: The intervals are the colorful horizontal lines presented in the following image. The Vertical black line represents the real value of α (in this case $\alpha = 7$). The red markers on the side denote which intervals *do not* contain the real α value. We can see that about 5 of the intervals do not contain 7. (When running the code sometimes we get more sometimes we get less). The y-values in the plot are meaningless of course.



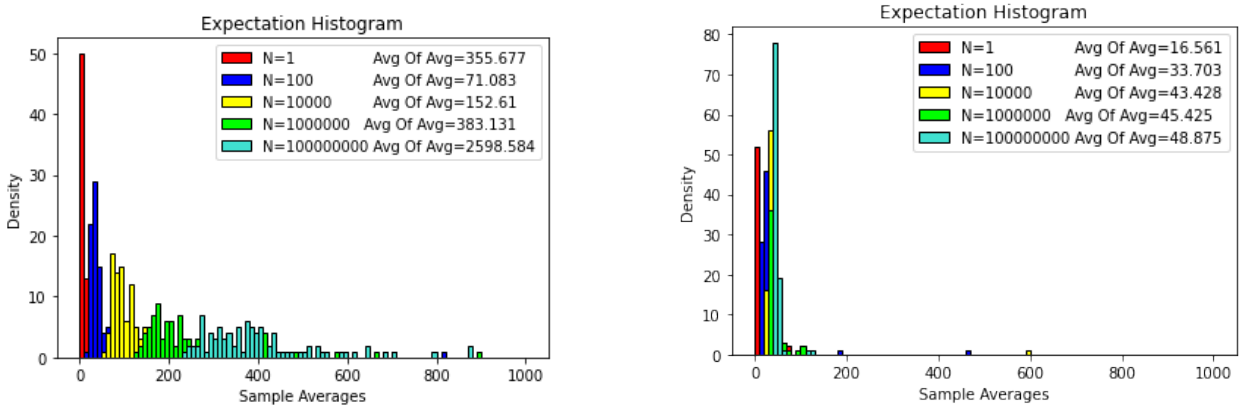
When looking at the metrics mentioned above we get the following results -

```
The average length of MLE based intervals is: 0.869342438275525
The standard deviation of the length of MLE based intervals is: 0.004052991420942892
The level of symmetry of the MLE based intervals is: 0.019106275266762916
The average length of pivot based intervals with 10^3 samples is: 0.8681816951455326
The standard deviation of the length of pivot based intervals with 10^3 samples is: 0.029033976196493422
The level of symmetry of the pivot based intervals with 10^3 samples is: 0.1787850834023662

The average length of pivot based intervals with 10^5 samples is: 0.08677452351518394
The standard deviation of the length of pivot based intervals with 10^5 samples is: 0.00026271659500458604
The level of symmetry of the pivot based intervals with 10^5 samples is: 0.01719071589635714
```

Initially this might imply that the MLE method is better than the pivot method because these metrics for the MLE method are equal to or better than the ones for the pivot method for the sample size of 10^3. The reason this is incorrect is because for the MLE method you need to sample 10^3 samples 10^4 times in order to get slightly better metrics than the pivot method with 10^3 samples, and if we even sample "only" 10^5 then the pivot method reaches significantly better results. This means the MLE method is much slower to calculate (running this in the code file we sent will show that), and since we can also sample many less samples for the pivot method there are memory advantages to the pivot method. All together this means the pivot method is better.

5. As is discussed in the code we use the sample average to estimate the mean of the sampling. The plots we got from the code are the following -



(a) $\alpha = 0.9$ (b) $\alpha = 1.1$

Figure 2: Histograms of sample averages for $\alpha = 0.9/1.1$

In the code file we present an explanation for why these results vary so much even though the α values are, at least to the human eye, pretty close. The explanation we provide for this is firstly based on understanding why in general the sample average is an estimate for the mean. Then there is a detailed explanation for why this is problematic for these values - for $\alpha = 0.9$ it doesn't work at all, and for $\alpha = 1.1$ it works but it's convergence doesn't "feel" very quick (where the suggested explanation considers the fact that the variance of the sampling is infinite). And so we get the obvious divergence between how the sample averages change between our α values. ■