Worksheet 10 - Maximum Likelihood Estimators and Simple Hypotheses

Shawn Kim

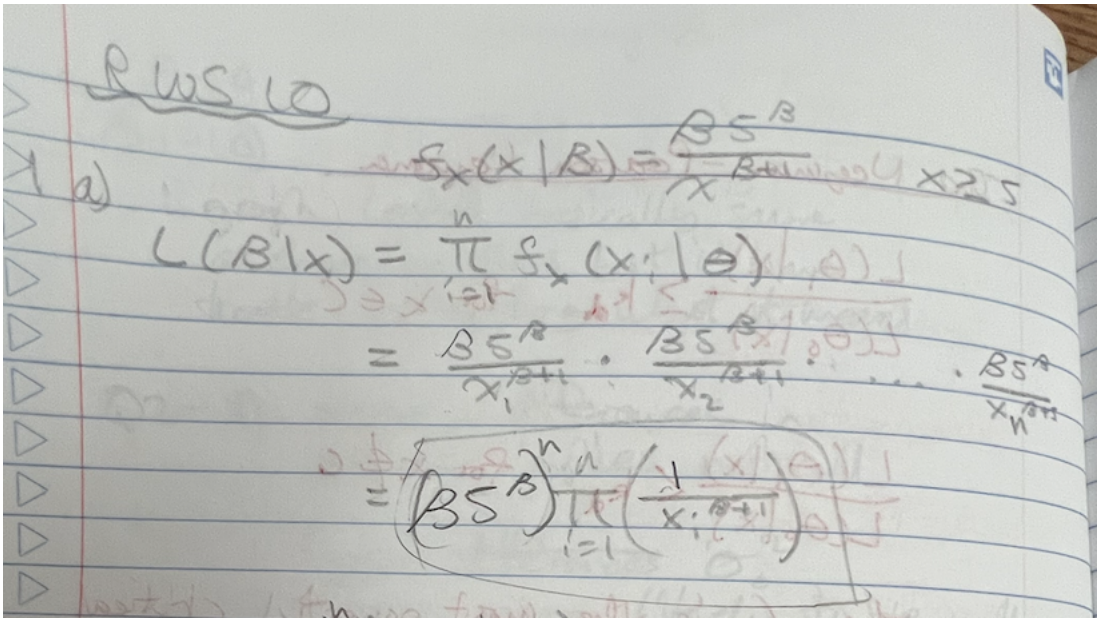
***Directions**: Please upload a PDF to Gradescope that includes both your written responses and corresponding R code inputs/outputs (if requested) for each problem.*

Problem 1. Loss of property for insurance purposes is sometimes modeled as a Pareto distribution. An insurance company offers two insurance policies. We will consider the claim amounts measured in thousands of dollars. The resulting density function for a minimum claim of 5 thousand dollars is

$$f_X(x|\beta) = \frac{\beta 5^\beta}{x^{\beta+1}}, \quad x \geq 5. \quad \text{with mean } \mu_X = \frac{5\beta}{\beta-1} \quad \text{and standard deviation } \sigma_X = \frac{5}{\beta-1}\sqrt{\frac{\beta}{\beta-2}}$$

Problem 1 Part a) Develop the likelihood function for n independent claims $x_1, x_2, ..., x_n$.

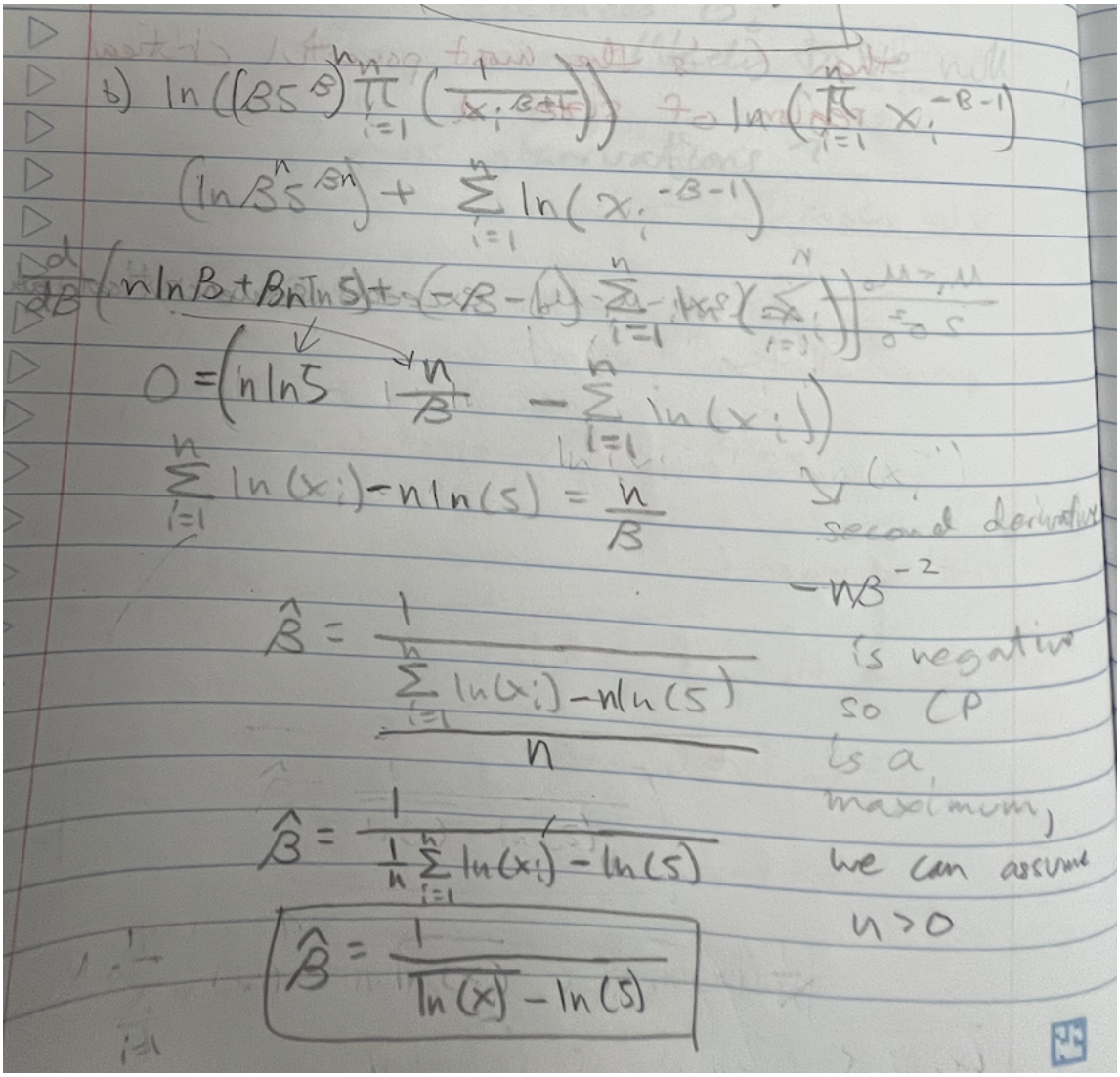
Hint: The likelihood function is very similar to the joint density function. In a joint density function: parameter β is known, and data $\mathbf{x}$ is unknown. In a likelihood function: parameter β is unknown, and the data $\mathbf{x}$ is known. So, we take the joint density function $f_X(x|\beta)$, choose to take the parameter β as unknown and the data $\mathbf{x}$ as known, arriving at the likelihood function, $L(\beta|\mathbf{x})$. Remember to simplify completely before moving on to part (b).



Problem 1 Part b) Find the maximum likelihood estimate $\hat{\beta}$ for β in terms of claims $x_1, x_2, \ldots, x_n$.

Hint: From part (a), find and simplify the log-likelihood function (recalling the properties of logarithms). Next, we optimize the log-likelihood function! This involves determining the score function (derivative with respect to the parameter), identifying the critical point, and confirming that this critical point is the maximum. (Note that a maximum of the log-likelihood function is known to also be a maximum of the likelihood function.)

Hint: You should get $\hat{\beta} = \frac{1}{\overline{\ln(x)} - \ln(5)}$, where $\overline{\ln(x)} = \frac{1}{n}\sum_{i=1}^{n} \ln(x_i)$, so $\hat{\beta}$ is a function of the claims data.



Problem 1 Part c) The claims under the insurance policy (i.e., the data) can be downloaded with the `R` command, `read.csv()`.

```
claims = read.csv("http://math.arizona.edu/~jwatkins/claims5.csv")
# NOTE: If you are having trouble accessing the CSV file via the URL, follow
# the instructions and uncomment the read.csv() command below. Be sure to
# comment out the read.csv() command above.

# Instructions: To change your working directory path to current file location,
```

```
# on the very top, click Session -> Set working directory -> To Source File
# Location OR Choose Directory. Be sure the CSV file and RMD file are in the
# same folder if using 'to Source File Location'

# claims = read.csv('claims.csv') # load CSV file directly from folder
```

★ Give the maximum likelihood estimate for these data.

```
claims = read.csv("http://math.arizona.edu/~jwatkins/claims5.csv")[, 1]


beta_hat = 1/(mean(log(claims)) - log(5))

beta_hat
```

```
## [1] 4.110196
```

Problem 1 Part d) Use the estimated parameter value from part (c) and the theoretical Pareto distribution to calculate the estimated mean and standard deviation of claims. Then compute the mean and standard deviation of the actual claims data. Produce a `data.frame` displaying your results.

```
# NOTE: Let MLEm be the mean of the Pareto distribution evaluated at beta_hat.
# Let MLEs be the standard deviation of the Pareto distribution evaluated at
# beta_hat. See the original problem statement for the necessary formulas

claims = read.csv("http://math.arizona.edu/~jwatkins/claims5.csv")[, 1]


(MLEm = (5 * beta_hat)/(beta_hat - 1))
```

```
## [1] 6.607616
```

```
(MLEs = (5/(beta_hat - 1)) * sqrt(beta_hat/(beta_hat - 2)))
```

```
## [1] 2.243635
```

```
# NOTE: Next, we compute the mean (DATAm) and SD (DATAs) of the data

(DATAm = mean(claims))
```

```
## [1] 6.613442
```

```
(DATAs = sd(claims))
```

```
## [1] 2.234332
```

```r
# NOTE: Now for the data.frame

means = c(MLEm, DATAm)  # collect the means
standevs = c(MLEs, DATAs)  # collect the standard deviations
data.frame(Mean = means, Standard_Dev = standevs, row.names = c("MLE", "data"))
```

```
##          Mean Standard_Dev
## MLE  6.607616     2.243635
## data 6.613442     2.234332
```

Finally, how well do the mean and standard deviation of the actual claims data match with that of the theoretical Pareto distribution?

according to the dataframe, the mean and standard deviations of the actual claims match closely with that of the Pareto distribution

Problem 2. *(From Session 19 Group Activity)*

The body temperature in degrees Fahrenheit of $n = 40$ randomly chosen healthy adults is measured. The standard deviation σ is known to be 0.68 degrees Fahrenheit. The sample mean for the measurements is $\bar{x} = 98.37$.

(Do Not Repeat Solutions) You already found a 99% confidence interval for the mean body temperature and explained its meaning.

Problem 2 Part a) Consider the simple hypothesis for the mean body temperature in degrees Fahrenheit

$$H_0 : \mu = 98.6 \quad \text{versus} \quad H_1 : \mu = 98.4.$$

Give the critical value for $\bar{x}$, the sample mean body temperature, when the significance level $\alpha = 0.05, 0.02$, and $\alpha = 0.01$. At what levels would you reject the null hypothesis?

```
(k_0.05 = -qnorm(0.95) * 0.68/sqrt(40) + 98.6)
```

```
## [1] 98.42315
```

```
(k_0.02 = -qnorm(0.98) * 0.68/sqrt(40) + 98.6)
```

```
## [1] 98.37919
```

```
(k_0.01 = -qnorm(0.99) * 0.68/sqrt(40) + 98.6)
```

```
## [1] 98.34988
```

you would reject the null hypothesis at the 0.05 significance level

Problem 2 Part b) If the number of healthy adults chosen were to increase to $n = 50$, while everything else remained unchanged, would the critical values increase, decrease, or stay the same when $\alpha = 0.05, 0.02, 0.01$? Justify your thinking.

the critcal values would increase if n were to increase to 50 observations because we can see from the equation that the value of k alpha is being subtracted by z alpha over sqrt of n and so the larger the n, the less is being subtracted from the null mean . graphically speaking, the more n observations we have, the more peaked the distributions become around their true means, thus decreasing the absolute value of the differnce bewteen the critical point and the true mean. we can also think of it like the more observations we have to be sure of the true mean of the null hypothesis, the more we can be sure that any sample mean of the same observation size that differs from the null mean is sufficient enough to reject the null hypothesis

Problem 2 Part c) Find the power of the test in part (b) for each significance level.

Hint: Begin by finding the z-score of the critical value, k_α, with respect to the alternative hypothesis. (Remember to account for the sample size when calculating the z-score.) Since the alternative distribution is to the left of the null distribution, the power is the probability $P\{Z \leq z\text{-score of } k_\alpha\}$ under the alternative distribution. You can check your answers using the Stats-Power Applet, https://digitalfirst.bfwpub.com/stats_applet/stats_applet_9_power.html

Hint: You have several methods to calculate the probability including `R`, your calculator, and the tables. I suggest practicing the calculations using your calculator or the tables since `R` is not a valid tool for the exams. Lastly, be sure to clearly state your inputs/outputs for whatever method you use.

```r
z_0.05 = qnorm(0.05, 98.6, 0.68/sqrt(50))
(power_0.05 = pnorm(z_0.05, 98.4, 0.68/sqrt(50)))
```

```
## [1] 0.6681724
```

```r
z_0.02 = qnorm(0.02, 98.6, 0.68/sqrt(50))
(power_0.02 = pnorm(z_0.02, 98.4, 0.68/sqrt(50)))
```

```
## [1] 0.5103621
```

```r
z_0.01 = qnorm(0.01, 98.6, 0.68/sqrt(50))
(power_0.01 = pnorm(z_0.01, 98.4, 0.68/sqrt(50)))
```

```
## [1] 0.4026004
```

Problem 2 Part d) On a single figure, plot the density of the distributions for the null and alternative hypothesis. On the plot, indicate the power for the case where $\alpha = 0.01$.

NOTE: To do this task, enter your code in the R chunk below by filling in the blanks denoted with `FILL IN` *and uncommenting all non-NOTE lines.*

```r
# NOTE: Begin by inputting the correct critical value from earlier

crit_value0.01 = qnorm(0.01, 98.6, 0.68/sqrt(50))

# NOTE: Now we plot the two density functions (which prefix is needed for
# density functions?). Be sure to define the correct mean and standard
# deviation for these normal distribution. Lastly, remember that the curve
# command requires that `x` be somewhere in the first entry, so FILL IN should
# be of the form: `prefix``family`(x, mean, sd) where the mean and sd are
# defined via the null distribution

curve(dnorm(x, 98.6, 0.68/sqrt(50)), from = 98.2, to = 98.8, xlab = "mean body temp",
    ylab = "density")

# NOTE: Now you need to call the necessary command to overlay plots and add the
# second plot for the alternative distribution

par(new = TRUE)
curve(dnorm(x, 98.4, 0.68/sqrt(50)), from = 98.2, to = 98.8, xlab = "", ylab = "",
    col = "red")

# NOTE: Next, we add a legend and draw a vertical segment at the critical value

legend("topright", legend = c("Dist. of H_0", "Dist. of H_1"), col = c("black", "red"),
    lty = 1, cex = 0.7)
segments(crit_value0.01, -0.5, crit_value0.01, 4, lty = 2)

# NOTE: Finally, we shade the power under the alternative distribution. Create
# a vector of x-values for the bottom of the shaded region
```

```r
x = seq(98, crit_value0.01, length = 100)

# NOTE: create a vector of y-values for the height of the shaded region this
# will be the same as one of the 'FILL IN' code for the density plots

y = dnorm(x, 98.4, 0.68/sqrt(50))


# NOTE: Use the polygon command! I'll leave this line in tact, since polygon
# can be tricky. You should try to understand the line below. Need help? See
# the Helpful Handout.

polygon(c(98, x, crit_value0.01), c(0, y, 0), col = "gray")
```