# STA380 Practice Problems for Quiz 4

These problems are not to be handed in, but they are for extra practice for students to be prepared for the quiz.

1. Consider the Pareto distribution:
$$f_X(x) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}}, \quad x > \beta.$$
Let $X_1, X_2, \ldots, X_n$ be a random sample from a Pareto distribution where $\beta = 2$, and we want to estimate $\alpha$. Note that the **method of moments estimator (MOME)** is:
$$\hat{\alpha}_{MOME} = \frac{\overline{Y}}{\overline{Y} - 2}$$
The solution can be found here. However, you do not need to recall how to find the method of moments estimator for this course.
   (a) Install the `extraDistr` package so we can simulate a random sample from the Pareto distribution, where $\beta = 2$ (`install.packages("extraDistr")`).
   (b) Use any size $n$ and any shape $\alpha$ to generate a random sample from the Pareto distribution. Use $m = 10^4$. (*Hint: read* `??extraDistr::Pareto`).
   (c) Compute a Monte Carlo estimate for $\hat{\alpha}_{MOME}$.
   (d) Derive the MLE for $\alpha$. (Yes, you need to remember how to compute the MLE for Unit 6.)
   (e) Compute a Monte Carlo estimate for $\hat{\alpha}_{MLE}$.
   (f) Determine which estimator is better: $\hat{\alpha}_{MOME}$ or $\hat{\alpha}_{MLE}$?

2. Let $X_1, X_2, \ldots, X_{10}$ be a random sample from the *Exponential*$(\theta)$ distribution, where $\theta$ is the *scale* parameter.
   (a) Use the NP Lemma to derive the **rejection region** (RR) for the uniformly most powerful test (UMP) of:
$$H_0 : \theta = 3 \quad \text{v.s.} \quad H_1 : \theta > 3$$
Let the significance level be $\alpha = 0.10$. *Hint:* for the final step, you are allowed to use `qexp()` or `qgamma()`. Also, for the exponential distribution, $\theta > 0$.
   (b) Use the following code to generate samples from the **null** distribution:
```
n <- 10
alpha <- 0.10
theta <- 3
m <- 10000
x <- matrix(rexp(n*m, rate = 1/theta), nrow = m)
```
Using the rejection region from part (a), compute the type-I error rate.
   (c) Assume we instead had the simple alternative hypothesis $H_1 : \theta = 4$. Using the rejection region from part (a), Compute the type-II error rate. Also, compute the power.
   (d) Why do you think the type-II error rate is high, and the power is low? What happens if you increase the value of $\theta$ for your alternative hypothesis?

Below are the solutions.

1. (a) The hint is the answer.
   (b) Below is an example for $n = 100$ and $\alpha = 3$.

```
n <- 100 # you can choose any value
m <- 10^4
alpha <- 3 # you can choose any value
y <- matrix(extraDistr::rpareto(n*m, a = alpha, b = 2), nrow = m)
```

   (c) `mome <- mean(rowMeans(y) / (rowMeans(y) - 2))`
   (d)

$$
\begin{aligned}
L(\alpha \,|\, \mathbf{X}) &= \alpha^n 2^{2n} \prod_{i=1}^{n} x_i^{-(\alpha+1)} \\
l(\alpha \,|\, \mathbf{X}) &= n\ln(\alpha) + \alpha n \ln(2) - (\alpha+1)\sum_{i=1}^{n} \ln(x_i) \\
\frac{d\, l(\alpha \,|\, \mathbf{X})}{d\alpha} &= \frac{n}{\alpha} + n\ln(2) - \sum_{i=1}^{n} \ln(x_i) \\
0 &\overset{set}{=} \frac{n}{\alpha} + n\ln(2) - \sum_{i=1}^{n} \ln(x_i) \\
\hat{\alpha}_{MLE} &= \frac{n}{\sum_{i=1}^{n} \ln(x_i) - n\ln(2)}
\end{aligned}
$$

   (e) `mle <- mean(n/(rowSums(log(y)) - n *log(2)))`
   (f) The MLE should perform better.

   In short, all of the `R` codes should be:

```
#install.packages("extraDistr")
n <- 100 # you can choose any value
m <- 10^4
alpha <- 3 # you can choose any value
y <- matrix(extraDistr::rpareto(n*m, a = alpha, b = 2), nrow = m)
mome <- mean(rowMeans(y) / (rowMeans(y) - 2))
mle <- mean(n/(rowSums(log(y)) - n *log(2)))
```

2. (a)

$$
L(\theta) = \frac{1}{\theta^{10}} \exp\left\{ -\frac{1}{\theta} \sum_{i=1}^{10} x_i \right\}
$$

   Thus,

$$
\begin{aligned}
\frac{L(\theta_0)}{L(\theta_a)} &= \frac{\frac{1}{\theta_0^{10}} \exp\left\{ -\frac{1}{\theta_0} \sum_{i=1}^{10} x_i \right\}}{\frac{1}{\theta_a^{10}} \exp\left\{ -\frac{1}{\theta_a} \sum_{i=1}^{10} x_i \right\}} \\
&= \left(\frac{\theta_a}{\theta_0}\right)^{10} \exp\left( \left( \frac{1}{\theta_a} - \frac{1}{\theta_0} \right) \sum_{i=1}^{10} x_i \right)
\end{aligned}
$$

   According to the NP lemma, the RR is determined by:

$$
\begin{aligned}
\frac{L(\theta_0)}{L(\theta_a)} &< k \\
\left(\frac{\theta_a}{\theta_0}\right)^{10} \exp\left( \left( \frac{1}{\theta_a} - \frac{1}{\theta_0} \right) \sum_{i=1}^{10} x_i \right) &< k \\
\exp\left( \left( \frac{1}{\theta_a} - \frac{1}{\theta_0} \right) \sum_{i=1}^{10} x_i \right) &< k_1 \\
\left( \frac{1}{\theta_a} - \frac{1}{\theta_0} \right) \sum_{i=1}^{10} x_i &< k_2
\end{aligned}
$$

Note that since we assumed that the null $H_0 : \theta = 3$ and the alternative hypothesis that $H_1 : \theta > 3$, we basically assume that:

$$\theta_a > \theta_0 \quad \Rightarrow \quad \frac{1}{\theta_a} < \frac{1}{\theta_0} \quad \Rightarrow \quad \frac{1}{\theta_a} - \frac{1}{\theta_0} < 0.$$

Continuing,

$$\left(\frac{1}{\theta_a} - \frac{1}{\theta_0}\right) \sum_{i=1}^{10} x_i < k_2$$
$$\sum_{i=1}^{10} x_i > k^*$$

We can use `qgamma()` to find the value for $k^*$ because:

$$\alpha = \mathbb{P}(RR \,|\, H_0 \text{ is true})$$
$$0.10 = \mathbb{P}\left(\sum_{i=1}^{10} x_i > k^* \,|\, \theta = 3\right)$$

Recall:

$$\sum_{i=1}^{10} x_i \sim Gamma(\alpha = 10, \beta = \theta)$$

So,

```
> k <- qgamma(0.1, shape = 10, scale = 3, lower.tail = FALSE)
> k
[1] 42.61797
```

Thus we let $k^* \approx 42.61797$. So, our rejection region is:

$$RR := \left\{\sum_{i=1}^{10} x_i > 42.61797\right\}$$

**Disclaimer:** admittedly I did skip a lot of steps, but that is because STA260 is a pre-requisite. If you need a refresher with more examples (and somewhat includes more steps), please see my STA260 tutorial 10 notes. I will not ask you to derive the RR on the exam (if there is a question about hypothesis testing, I will give you the rejection region). However, there will be a question on the quiz, which is completely open book.

(b)
```
n <- 10
alpha <- 0.10
theta <- 3
m <- 10000
# generating from the null distribution
x <- matrix(rexp(n*m, rate = 1/theta), nrow = m)
# computing the test statistic
sumx <- rowSums(x)
# computing the critical value
k <- qgamma(alpha, shape = 10, scale = theta, lower.tail = FALSE)
# reject dependent on the rejection region
type_1_err <- mean(sumx > k)
type_1_err
```

(c) Recall the following:

```
n <- 10
alpha <- 0.10
theta_1 <- 4
theta_0 <- 3
m <- 10000
x <- matrix(rexp(n*m, rate = 1/theta_1), nrow = m)
# computing the test statistic
sumx <- rowSums(x)
# computing the critical value
```

```
k <- qgamma(alpha, shape = 10, scale = theta_0, lower.tail = FALSE)
# accept dependent on the rejection region
type_2_err <- mean(sumx < k)
type_2_err
power = 1 - type_2_err
```

(d) Two issues: small sample size, and 4 is somewhat close to 3. If you increase the value for $\theta$ from the alternative hypothesis, your type-II error will decrease and the power will increase. This is a classical example as to why simple hypothesis testing can easily be manipulated and thus should be **eradicated**. (I am only partially joking.)