Homework 1: Written Part

STAT 343: Mathematical Statistics

SOLUTIONS

Details

How to Write Up

The written part of this assignment can be either typeset using latex or hand written.

Grading

5% of your grade on this assignment is for turning in something legible and organized.

An additional 15% of your grade is for completion. A quick pass will be made to ensure that you've made a reasonable attempt at all problems.

Across both the written part and the R part, in the range of 1 to 3 problems will be graded more carefully for correctness. In grading these problems, an emphasis will be placed on full explanations of your thought process. You don't need to write more than a few sentences for any given problem, but you should write complete sentences! Understanding and explaining the reasons behind what you are doing is at least as important as solving the problems correctly.

Solutions to all problems will be provided.

Collaboration

You are allowed to work with others on this assignment, but you must complete and submit your own write up. You should not copy large blocks of code or written text from another student.

Sources

You may refer to our text, Wikipedia, and other online sources. All sources you refer to must be cited in the space I have provided at the end of this problem set.

Problem I: χ^2, t, and F distributions

Note, for (1) and (2), you only need to justify your answers using properties of these distributions. You DO NOT need to derive anything using moment generating functions.

(1) Suppose that $Z_1, Z_2, Z_3, Z_4 \overset{\text{i.i.d.}}{\sim}$ Normal$(0, 1)$, and $Y = \frac{Z_1}{\sqrt{(Z_2^2+Z_3^2+Z_4^2)/3}}$. What is the distribution of Y? What is the distribution of Y^2? Justify your answers briefly.

$Y = \frac{Z_1}{\sqrt{(Z_2^2+Z_3^2+Z_4^2)/3}} \sim t_3$ for the following reasons:

- Z_1 is a standard normal.
- Z_2^2, Z_3^2, Z_4^2 are each χ^2_1, because they are squared standard normal random variables.
- $Z_2^2 + Z_3^2 + Z_4^2 \sim \chi^2_3$, because the sum of three independent chi-squared random variables with 1 degree of freedom is also chi-squared, and we sum the degrees of freedom.

- The numerator and denominator are independent because Z_1 is independent of Z_2, Z_3, Z_4 (and thus of $Z_2^2+Z_3^2+Z_4^2$).
- So, since $Z_1 \sim Normal(0,1)$ and $Z_2^2+Z_3^2+Z_4^2 \sim \chi_3^2$ and they are independent, then we get the distribution result noted above (t_3).

$Y^2 \sim F_{1,3}$ for the following reasons:

- Rewrite $Y^2 = \left[\frac{Z_1}{\sqrt{(Z_2^2+Z_3^2+Z_4^2)/3}}\right]^2 = \frac{Z_1^2/1}{(Z_2^2+Z_3^2+Z_4^2)/3}$.
- $Z_1^2 \sim \chi_1^2$ and $Z_2^2+Z_3^2+Z_4^2 \sim \chi_3^2$ as established in (1).
- Z_1^2 and $Z_2^2+Z_3^2+Z_4^2$ are independent.
- Thus, $Y^2 = \frac{Z_1^2/1}{(Z_2^2+Z_3^2+Z_4^2)/3}$ has the F distribution noted above.

(2) Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim}$ Normal(μ_X, σ^2) and $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim}$ Normal(μ_Y, σ^2), with all X's and Y's independent. Note that the two distributions have different means but the same variance. Show how you can use the F distribution to find $P(S_X^2/S_Y^2 > 2)$, where S_X^2 is the sample variance of $X_1, \ldots, X_n$ and S_Y^2 is the sample variance of $Y_1, \ldots, Y_m$. We know from the t, F, chi-squared notes that:

(1) $\frac{(n-1)S_X^2}{\sigma^2} \sim \chi_{n-1}^2$
(2) $\frac{(m-1)S_Y^2}{\sigma^2} \sim \chi_{m-1}^2$
(3) All X's and Y's are independent, so functions of them (S_X^2 and S_Y^2) are also independent

Using these pieces of information, we know

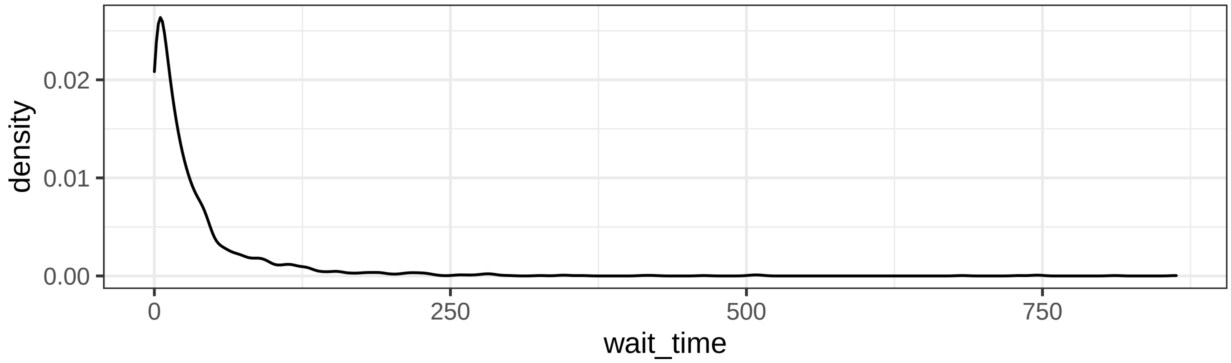
$$\frac{S_X^2}{S_Y^2} = \frac{[(n-1)S_X^2/\sigma^2]/(n-1)}{[(m-1)S_Y^2/\sigma^2]/(m-1)} \sim F_{(n-1),(m-1)}$$

Once we know this, we can use this fact to find the probability $P(S_X^2/S_Y^2 > 2)$.

Problem II: Emergency Department Waiting Times

The National Center for Health Statistics, a division within the U.S. Centers for Disease Control, conducts a nationally representative survey of hospitals each year to track the waiting times for emergency room visits (that is, how much time passed between when a patient arrived at the hospital and when they were seen by a doctor or registered nurse). In this problem, we will model the distribution of waiting times for 1874 emergency department visits from 2012. The output below shows a plot of the data.

```
ggplot(data = er_visits, mapping = aes(x = wait_time)) +
  geom_density() +
  theme_bw()
```



An exponential distribution is often used to model waiting times. Suppose we adopt the data model

$$X_i \overset{\text{i.i.d.}}{\sim} \text{Exponential}(\theta),\ i = 1, \dots, n,$$

where X_i is the waiting time for visit number i. In our data set we have observed values $x_1, \dots, x_n$, where $n = 1874$.

Here are some facts about the exponential distribution (note that there are two common parameterizations of the exponential distribution; please work with the definitions and properties stated below for this problem):

parameter	$\theta > 0$: rate parameter
p.f.	$f_{X\|\Theta}(x\|\theta) = \theta e^{-x\theta}$ on the support $x \geq 0$
Mean	$\frac{1}{\theta}$
Variance	$\frac{1}{\theta^2}$

(1) Find the maximum likelihood estimator of θ. Since $X_1, ..., X_n$ are i.i.d. Exponential(θ), the likelihood is

$$\begin{aligned} \mathcal{L}(\theta|X_1, ..., X_n) &= \prod_{i=1}^{n} \theta e^{-X_i\theta} \text{ by independence} \\ &= \theta^n e^{-\theta \sum_{i=1}^{n} X_i} \text{ by multiplication and properties of exponents} \end{aligned}$$

for $X_i \in (0, \infty)\ \forall i$ and $\theta > 0$ (and 0 otherwise).

Now, let's find the log likelihood:

$$\ell(\theta|X_1, ..., X_n) = n \log\theta - \theta \sum_{i=1}^{n} X_i$$

for $X_i \in (0, \infty)\ \forall i$ and $\theta > 0$.

Next, we find the critical point by setting the first derivative with respect to θ equal to 0, and solving for θ:

$$\begin{aligned} 0 &= \frac{d}{d\theta}\left[n \log\theta - \theta \sum_{i=1}^{n} X_i\right] \\ &= \frac{n}{\theta} - \sum_{i=1}^{n} X_i \\ \Rightarrow \hat{\theta} &= \left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]^{-1} = \bar{X}^{-1} \end{aligned}$$

(The critical point is the inverse of the sample mean.)

We need the second derivative test to verify the critical point is a global maximum:

(WTS: $\frac{d^2}{d^2\theta}\ell(\theta|X_1, ..., X_n) < 0\ \forall\theta$)

$$\begin{aligned} \frac{d^2}{d^2\theta}\ell(\theta|X_1, ..., X_n) &= \frac{d}{d\theta}\left[\frac{n}{\theta} - \sum_{i=1}^{n} X_i\right] \\ &= -\frac{n}{\theta^2} \\ &< 0\ \forall\theta \end{aligned}$$

Therefore, the maximum likelihood estimator is $\hat{\theta}^{MLE} = \left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]^{-1}$.

(Grader Note: Also give full credit if the student plugs in $n = 1874$.)

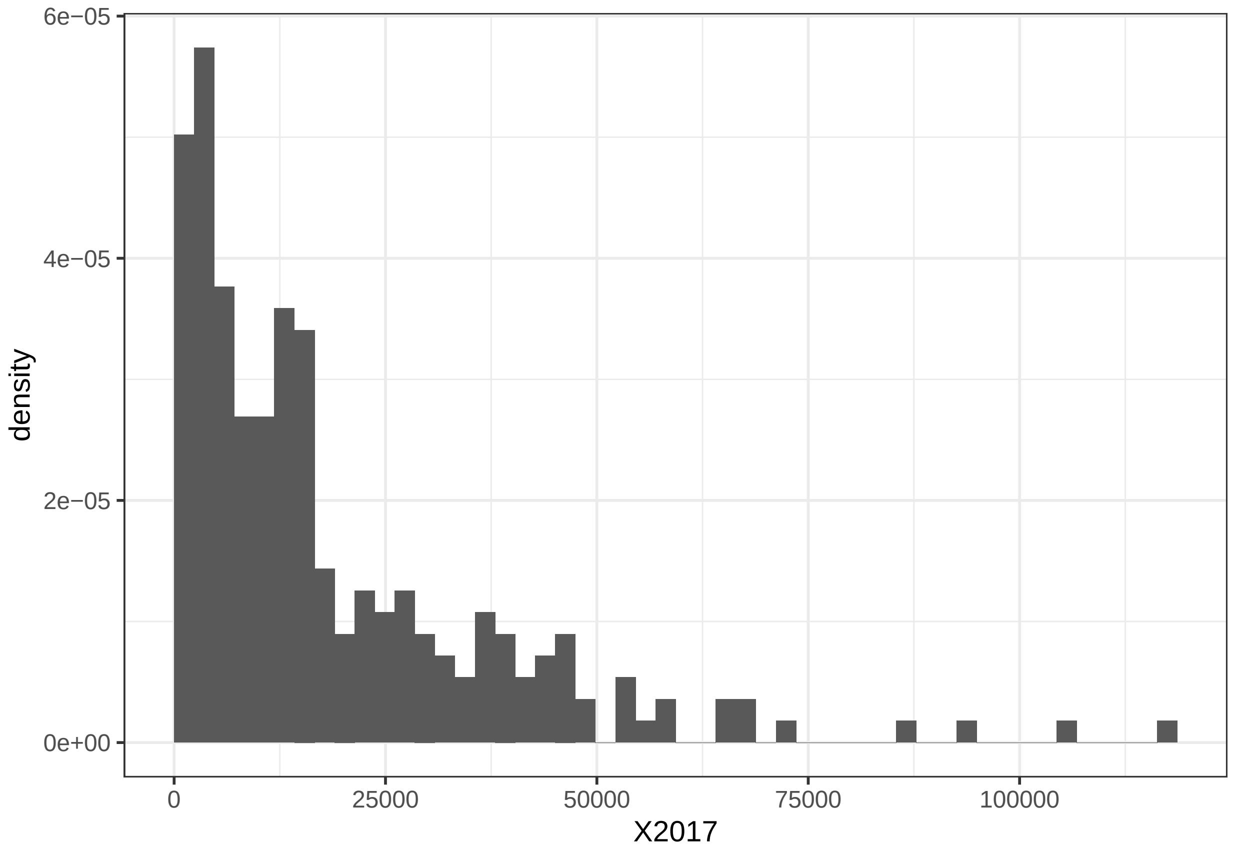
(2) Is your result from part (1) a random variable or a number? If it is a random variable, explain why it is random. If it is a number, explain why it is not random. (Your answer should be more detailed than "because that's the definition of an estimator.") \

The result from Part (1) is a random variable. It is random because one could get a different sample mean depending on the sample.

Problem III: Per Capita GDP

The code below reads in and plots a data set with measurements of per capita GDP at purchasing power parity as of 2017 for 235 countries, measured in inflation-adjusted 2011 international dollars; these data are from the World Bank, here: https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.KD. Per capita GDP can be roughly interpreted as the amount of income generated in a country in one year divided by the number of people living in that country. The purchasing power parity adjustment attempts to adjust GDP to account for differences in cost of living in different countries.

```r
gdp <- read.csv("https://marievozanne.github.io/stat343/data/gdp.csv")
gdp <- gdp %>%
  filter(!is.na(X2017))
ggplot(data = gdp, mapping = aes(x = X2017)) +
  geom_histogram(mapping = aes(y = ..density..), boundary = 0, bins = 50) +
  theme_bw()
```



A lognormal distribution is often used to model non-negative variables that are skewed right, like incomes. In the written part of this assignment you will find the maximum likelihood estimator for the parameters of a lognormal distribution, and in the R part of the assignment you will fit the model to this data set.

For the purpose of this assignment, let's assume that the per capita GDP of different countries in a given year can be modeled as independent, identically distributed random variables (this is not actually reasonable, but may be good enough if our goal is to describe the distribution of values for per capita GDP across different countries).

Let's adopt the model $X_i \overset{\text{i.i.d.}}{\sim} \text{lognormal}(\mu, \sigma)$, $i = 1, \dots, n$.

The pdf of a lognormal distribution is given by $f(x|\mu,\sigma) = x^{-1}(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\frac{\{\log(x)-\mu\}^2}{\sigma^2}\right]$

Find the maximum likelihood estimators of μ and σ (in terms of x). For this problem, you do not have to check second-order conditions to verify that you have found a global maximum of the log-likelihood function. The likelihood function is

$$\begin{aligned}
\mathcal{L}(\mu,\sigma|X_1,...,X_n) &= \prod_{i=1}^{n} X_i^{-1}(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\frac{\{\log(X_i)-\mu\}^2}{\sigma^2}\right] \\
&= (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{\{\log(X_i)-\mu\}^2}{\sigma^2}\right]\prod_{i=1}^{n}X_i^{-1}
\end{aligned}$$

The log-likelihood function is

$$\ell(\mu,\sigma|X_1,...,X_n) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2}\sum_{i=1}^{n}\frac{\{\log(X_i)-\mu\}^2}{\sigma^2} + \sum_{i=1}^{n}\log\left(X_i^{-1}\right)$$

There may be minor differences in how you chose to simplify your expression.

Solving for the critical point for μ:

$$\begin{aligned}
0 &= \frac{\partial}{\partial\mu}\ell(\mu,\sigma|X_1,...,X_n) \\
&= \frac{\partial}{\partial\mu}\left[-\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2}\sum_{i=1}^{n}\frac{\{\log(X_i)-\mu\}^2}{\sigma^2} + \sum_{i=1}^{n}\log\left(X_i^{-1}\right)\right] \\
&= 0 - 0 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial}{\partial\mu}\{\log(X_i)-\mu\}^2 + 0 \\
&= -\frac{1}{2\sigma^2}\times 2\sum_{i=1}^{n}(\log(X_i)-\mu)(-1) \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(\log(X_i)-\mu) \\
\Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}\log(X_i) &= \frac{1}{\sigma^2}(n\mu) \\
\Rightarrow \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n}\log(X_i)
\end{aligned}$$

Solving for the critical point for σ:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \sigma} \ell(\mu, \sigma | X_1, ..., X_n) \\
&= \frac{\partial}{\partial \sigma} \left[-\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2}\sum_{i=1}^{n} \frac{\{\log(X_i) - \mu\}^2}{\sigma^2} + \sum_{i=1}^{n} \log\left(X_i^{-1}\right) \right] \\
&= -\frac{n}{\sigma} - \frac{1}{2}\sum_{i=1}^{n} \{\log(X_i) - \mu\}^2 \times (-2)\left(\frac{1}{\sigma^3}\right) \\
&= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n} \{\log(X_i) - \mu\}^2 \\
\Rightarrow & \frac{n}{\sigma} = \frac{1}{\sigma^3}\sum_{i=1}^{n} \{\log(X_i) - \mu\}^2 \\
\Rightarrow & n = \frac{1}{\sigma^2}\sum_{i=1}^{n} \{\log(X_i) - \mu\}^2 \\
\Rightarrow & \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \{\log(X_i) - \mu\}^2 \\
\Rightarrow & \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \{\log(X_i) - \hat{\mu}\}^2}
\end{aligned}
$$

Since we already have an estimate for μ, we can plug that in, so

$$
\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left\{ \log(X_i) - \frac{1}{n}\sum_{i=1}^{n} \log(X_i) \right\}^2}.
$$

Since the problem said we did not have to verify a maximum for these critical points, we are done!