Question 1

The goal of this question is to prove Theorem 1.6.13: for a random variable X with distribution F_X and median m, then for any real value a,

$$\min_a \mathrm{E}\left(\,|X-a|\,\right) = \mathrm{E}\left(\,|X-m|\,\right).$$

While it is possible to prove certain special cases (e.g. assuming X is continuous and has a p.d.f.) relatively easily, this question breaks down a proof for the general case.

Consider the function $G(X) = |X-a| - |X-m|$. Then, assuming $a \geq m$,

(a) Show that $G(X) = a - m$ when $X \leq m$, and $G(X) \geq -(a-m)$ when $X > m$.

(b) Using the indicator function $\mathbb{I}(X \leq m)$, show that $\mathrm{E}(G(X)\,\mathbb{I}(X \leq m)) = (a-m)\mathrm{P}(X \leq m)$.

(c) Using the indicator function $\mathbb{I}(X > m)$, show that $\mathrm{E}(G(X)\,\mathbb{I}(X > m)) \geq -(a-m)\mathrm{P}(X > m)$

(d) Using Parts (b) and (c), show that $\mathrm{E}(G(X)) \geq (a-m)\left[\mathrm{P}(X \leq m) - \mathrm{P}(X > m)\right]$.

(e) Recalling from the definition of the median that $\mathrm{P}(X \leq m) \geq \frac{1}{2}$, show that $\mathrm{E}(G(X)) \geq 0$.

(f) Conclude that the theorem is true if we can assume $a \geq m$.

(g) Show the theorem is true, no matter the value of a.

Solution to Question 1

Part (a):

To prove the theorem one needs to show, for a random variable X with a median m, that for any value a, $\mathrm{E}(|X-a|) \geq \mathrm{E}(|X-m|)$. This is equivalent to showing that $\mathrm{E}(|X-a| - |X-m|) \geq 0$, which leads us to define the function

$$G(X) = |X-a| - |X-m|.$$

We also start by assuming that $a \geq m$. Then in order to investigate the value of $G(X)$, we need to look at the three cases:

1. $X \leq m (\leq a)$,
2. $m \leq X \leq a$,
3. $X \geq a (\geq m)$.

Case 1: For $X \leq m \leq a$:

$$\begin{aligned} |X-a| - |X-m| &= -(X-a) - (-(X-m)) \\ &= a - m \end{aligned}$$

Case 2: For $m \leq X \leq a$, the situation is slightly different:

$$\begin{aligned} |X-a| - |X-m| &= -(X-a) - (X-m) \\ &= -X + a - X + m \\ &= -2X + a + m, \end{aligned}$$

and then since $X \leq a \Rightarrow -2X \geq 2a$,

$$|X-a| - |X-m| = -2X + a + m \geq -2a + a + m = -(a-m).$$

Case 3: For $X \geq a\,(\geq m)$:

$$\begin{aligned} |X-a| - |X-m| &= (X-a) - (X-m) \\ &= X - a - X + m \\ &= -(a-m). \end{aligned}$$

We can summarise these three cases in two cases:

$$\begin{aligned} X \leq m: &\qquad G(X) = |X-a| - |X-m| = a-m, \\ X > m: &\qquad G(X) = |X-a| - |X-m| \geq -(a-m). \end{aligned}$$

Note that $a - m \geq 0$, and that we consider the event $\{X > m\}$ rather than $\{X \geq m\}$, so that $\{X \leq m\}$ and $\{X > m\}$ are complementary events.

Part (b):

Now we use the indicator variable $\mathbb{I}_{\{X \leq m\}} = \mathbb{I}(X \leq m)$:

$$\begin{aligned} &\text{if } X \leq m,\ G(X) = a - m \\ &\qquad \Rightarrow G(X)\,\mathbb{I}(X \leq m) = (a-m)\mathbb{I}(X \leq m) \\ &\qquad \Rightarrow \mathrm{E}\left(G(X)\,\mathbb{I}(X \leq m)\right) = \mathrm{E}\left((a-m)\mathbb{I}(X \leq m)\right) \\ &\qquad \Rightarrow \mathrm{E}\left(G(X)\,\mathbb{I}(X \leq m)\right) = (a-m)\mathrm{E}\left(\mathbb{I}(X \leq m)\right) \\ &\qquad \Rightarrow \mathrm{E}\left(G(X)\,\mathbb{I}(X \leq m)\right) = (a-m)\mathrm{P}\left(X \leq m\right), \end{aligned} \tag{1}$$

where we recall from the last problem sheet that for any event A, $\mathbb{I}(A) = \mathrm{P}(A)$.

Note that the second line,

$$G(X)\,\mathbb{I}(X \leq m) = (a-m)\mathbb{I}(X \leq m)$$

no longer needs the condition that $X \leq m$. This line shows that if $X \leq m$ then $G(X) = (a-m)$, however if $X > m$, then this line becomes $0 = 0$, which is also a true statement. The indicator function provides a nice way to include a condition on a random variable in an equation.

Part (c):

Similarly, using the indicator variable $\mathbb{I}(X > m)$, we can show that

$$\begin{aligned} &\text{if } X > m,\ G(X) \geq -(a-m) \\ &\qquad \Rightarrow G(X)\,\mathbb{I}(X > m) \geq -(a-m)\mathbb{I}(X > m) \\ &\qquad \Rightarrow \mathrm{E}\left(G(X)\,\mathbb{I}(X > m)\right) \geq \mathrm{E}\left(-(a-m)\mathbb{I}(X > m)\right) \\ &\qquad \Rightarrow \mathrm{E}\left(G(X)\,\mathbb{I}(X > m)\right) \geq -(a-m)\mathrm{E}\left(\mathbb{I}(X > m)\right) \\ &\qquad \Rightarrow \mathrm{E}\left(G(X)\,\mathbb{I}(X > m)\right) \geq -(a-m)\mathrm{P}\left(X > m\right) \end{aligned} \tag{2}$$

Part (d):

Noticing that we use the indicator variables to partition 1:

$$\begin{aligned} 1 &= \mathbb{I}(X \leq m) + \mathbb{I}(X > m) \\ \Rightarrow G(X) &= G(X)\mathbb{I}(X \leq m) + G(X)\mathbb{I}(X > m) \\ \Rightarrow \mathrm{E}[G(X)] &= \mathrm{E}[G(X)\mathbb{I}(X \leq m)] + \mathrm{E}[G(X)\mathbb{I}(X > m)] \end{aligned} \tag{3}$$

where the last line follows from the linearity of expectation. Then

$$\begin{aligned} &\mathrm{E}[G(X)] = (a-m)\mathrm{P}\left(X \leq m\right) + \mathrm{E}[G(X)\mathbb{I}(X > m)] && \text{(using Part (b), Equations (2) and (1))} \\ \Rightarrow\ &\mathrm{E}[G(X)] \geq (a-m)\mathrm{P}\left(X \leq m\right) - (a-m)\mathrm{P}\left(X > m\right) && \text{(using Part (c), using Equation(3))} \\ \Rightarrow\ &\mathrm{E}[G(X)] \geq (a-m)\left[\mathrm{P}\left(X \leq m\right) - \mathrm{P}\left(X > m\right)\right]. \end{aligned}$$

Part (e):

From the definition of a median,

$$\begin{aligned} &\mathrm{P}(X \leq m) \geq \tfrac{1}{2} \\ \Rightarrow &2\mathrm{P}(X \leq m) - 1 \geq 0 \end{aligned}$$

Now, any random variable X,

$$\begin{aligned} &\mathrm{P}(X \leq m) + \mathrm{P}(X > m) = 1 \\ \Rightarrow &\mathrm{P}(X > m) = 1 - \mathrm{P}(X \leq m). \end{aligned}$$

Then, starting with Part (d), and noting that $a - m \geq 0$,:

$$\begin{aligned} \mathrm{E}[G(X)] &\geq (a-m)\left[\mathrm{P}(X \leq m) - \mathrm{P}(X > m)\right] \\ &= (a-m)\left[\mathrm{P}(X \leq m) - (1 - \mathrm{P}(X \leq m))\right] \\ &= (a-m)\left[2\mathrm{P}(X \leq m) - 1\right] \\ &\geq (a-m) \cdot 0 \\ &\geq 0. \end{aligned}$$

Part (f):

Starting with Part (e),

$$\begin{aligned} &\mathrm{E}(G(X)) \geq 0 \\ &\mathrm{E}(|X-a| - |X-m|) \geq 0 \\ \Rightarrow &\mathrm{E}(|X-a|) - \mathrm{E}(|X-m|) \geq 0 \\ \Rightarrow &\mathrm{E}(|X-a|) \geq \mathrm{E}(|X-m|), \end{aligned}$$

using the linearity of the expectation. This shows that m minimises the function $\mathrm{E}(|X-a|)$ (as a function of a), for the case $a \geq m$.

Part (g):

A similar argument to that used in Parts (a) to (d) can be used for the case $a \leq m$, which proves the result.

(Again defining $G(X) = |X-a| - |X-m|$, we consider the cases:

$$\begin{aligned} (1)\ X \leq a (\leq m) \Rightarrow |X-a| - |X-m| &= -(X-a) - (-(X-m)) \\ &= a - m \\ &= -(m-a) \\ (2)\ a \leq X \leq m \Rightarrow |X-a| - |X-m| &= (X-a) - (-(X-m)) \\ &= 2X - a - m \\ &\geq 2a - a - m = -(m-a) \\ (3)\ a \leq m \leq X \Rightarrow |X-a| - |X-m| &= (X-a) - (X-m) \\ &= m - a \end{aligned}$$

which can be summarised by

$$\begin{aligned} X < m: &\quad G(X) = |X-a| - |X-m| \geq -(m-a) \\ X \geq m: &\quad G(X) = |X-a| - |X-m| = m - a \end{aligned}$$

and from here the result follows similarly to the first case.)

Alternative approach to (g):

An alternative approach is to use symmetry from the first case. Note that we have proved:

$$\text{If } X \text{ is a random variable with median } m, \text{ and } a \geq m, \text{ then } \mathrm{E}(|X-a|) \geq \mathrm{E}(|X-m|)\ . \tag{4}$$

We reparametrise this result by setting $Y = X - m$, and noting that Y has median 0. We also set $b = a - m \geq 0$. Then

$$\begin{aligned} &\mathrm{E}(|X-a|) \geq \mathrm{E}(|X-m|) \\ \Rightarrow &\mathrm{E}(|Y+m-a|) \geq \mathrm{E}(|Y|) \\ \Rightarrow &\mathrm{E}(|Y-b|) \geq \mathrm{E}(|Y|)\,, \end{aligned}$$

which shows that Result (4) is equivalent to

$$\text{If } Y \text{ is a random variable with median 0, and } b \geq 0, \text{ then } \mathrm{E}(|Y-b|) \geq \mathrm{E}(|Y|)\ . \tag{5}$$

We would like to show this result also holds when the constant is less than or equal to zero, i.e. for $c \leq 0$, $\mathrm{E}(|Y-c|) \geq \mathrm{E}(|Y|)$. This is where symmetry is used; suppose that $c \leq 0$, then set $b = -c \geq 0$ and set $Z = -Y$. Note that Z also has 0 as a median. Then, using $|-W| = |W|$ for any W,

$$\begin{aligned} \mathrm{E}(|Y-c|) = \mathrm{E}(|-Z+b|) = \mathrm{E}(|Z-b|) &\geq \mathrm{E}(|Z|) && \text{(using Result (5), with } Z) \\ &= \mathrm{E}(|-Y|) \\ &= \mathrm{E}(|Y|) \\ \Rightarrow \mathrm{E}(|Y-c|) \geq \mathrm{E}(|Y|)\,. & \end{aligned}$$

Some comments The above proof seems very long, but only because all the steps have been carefully written out, since this is perhaps the first time we have used indicator variables in this way. In hindsight, the proof itself is actually quite easy; (1) we want to prove something is a minimum, so (2) we rewrite it as a function $G(X)$ that we want to show has some simpler property (nonnegative expectation), (3) we check a few cases to see the value of this function over the whole real line, (4) it turns out there are only really two special regions worth considering ($X \leq m$ and $X > m$), (5) use the indicator function to split the definition of $G(X)$ over the two regions, (6) some straightforward algebra, and we have the result.

That being said, this proof is only easy if you know the technique—if one has never seen this sort of technique before, it requires a lot of thought!

Another approach would be to reparametrise the problem from the start, and rather prove $\mathrm{E}(|Y-b|) \geq \mathrm{E}(|Y|)$ for Y a random variable with median 0 and $b \geq 0$ (and then use the same symmetry argument for $b \leq 0$). This would perhaps be more efficient since it would cut down the three cases to consider in Part (a) to only two cases.

Question 2 (R question)

In Question 3 you will create an R Markdown document. This question helps you to ensure that R Markdown is installed.

Open R Studio and then click on the '+' icon in the top-left corner which gives you options to create a new file. Select the option 'R Markdown'. RStudio may then display a message saying you need certain packages to be installed in order to create an R Markdown document, and ask if you want to install these packages; say yes. After a while, the packages will be installed and you can move on to the next question.

Alternative The above is perhaps the easiest way to get the packages installed. There is also a more 'manual' method: open R Studio, and then in the **in the R console** run the command

```
library(rmarkdown)
```

If the command runs without throwing an error, then great - R Markdown is installed! If there is an error, then **in the R console** run

```
install.packages("rmarkdown")
```

which will install the `rmarkdown` package and a few other packages that are necessary. Note that there is a way to ensure that R packages are installed to a specific directory by specifying the `lib` path in the `install.packages` command, but then you will need to make sure that this path is added to `.libPaths()` whenever you start R.

Solution to Question 2 (R Question)

No solution, just follow the steps.

Question 3 (R question)

This question is to ensure you can generate an R Markdown report using R Studio.

(a) Open RStudio and then open the file `example_rmarkdown_to_pdf.Rmd` in RStudio. The file is available to download on Blackboard.

(b) Click the 'Knit' button to generate a PDF version of this file. If you cannot see a 'Knit' icon, look for a 'Preview' icon. Or click on the drop-down arrow next to 'Knit' (or 'Preview') and select 'Knit to PDF'.

(c) A PDF should open showing a scatterplot. There are two possible things that can go wrong in this step, though.

- If the PDF opens, but the scatterplot does not appear, in RStudio click on the drop-down arrow on 'Run' command (top-right corner of editor window) and select 'Run All' and 're-knit' (which is like re-compiling).
- If the PDF does not open at all, and you see an error in the console, then please see the 'Troubleshooting' section below.

(d) Modify the 'subtitle' field at the top of your Rmd file to have your full name and CID number.

(e) Edit the Rmd file to fix the expression for the normal probability density function. This is an example of LaTeX code for creating equations; although you may not have seen it before, try to match the code with the output and change the numbers accordingly. After editing, click 'Knit' again to regenerate the PDF

(f) Edit the Rmd file to plot a histogram instead of a scatterplot.

(g) Increase the number of observations generated from 100 to 10000. Re-knit the document. If the changes do not take effect, re-knit the document a second time.

(h) For the histogram, set the parameter `freq` to `FALSE`, i.e. `hist(z, freq=FALSE)`, and replot the histogram (re-knit the document).

(i) Plot the probability density function for the appropriate normal distribution over the histogram. You will need to use the `lines` command, and set the parameters for the normal distribution's mean and standard deviation inside the `dnorm` command correctly.

(j) Save the PDF of the R Markdown document to your computer.

Solution to Question 3 (R Question)

For the solution, see the file `solution_Q3_PS10.Rmd`. The output of this file is shown on the next pages.

MATH40005 Problem Sheet 10 Question 3

Ronald A. Fisher, CID: 12345678

Introduction

The probability density function of a normal distribution with mean μ and variance σ^2 is

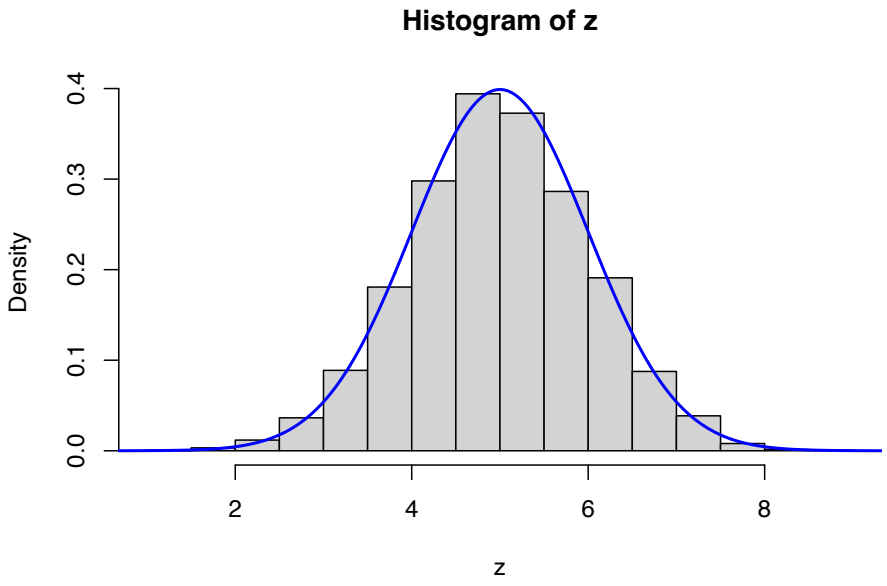
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

A histogram

```
# set the mean and standard deviation for a normal distribution,
# choose your own parameter values
mu <- 5
sigma <- 1

# generate observations followi ng a normal distribution with those parameter values
set.seed(1)
z <- rnorm(n=10000, mean=mu, sd=sigma)

# plotting the data, with a histogram and overlaying a density
hist(z, freq=FALSE)
k <- 5
x <- seq(from=mu-k*sigma, to=mu+k*sigma, by=0.01)
y <- dnorm(x, mean=mu, sd=sigma)
lines(x=x, y=y, type='l', lwd=2, col="blue")
```


Histogram of z
Density
0.0
0.1
0.2
0.3
0.4
2
4
6
8
z