**利用R來做出下列圖表：**

**Table 2.2**

The values needed to calculate the ECDF for the grouped pizza delivery time data in Example 2.2.3

Delivery time j ej−1 ej nj fj F(ej)

[0;10] 1 0 10 0 0.0000 0.0000

(10;15] 2 10 15 3 0.0024 0.0024

(15;20] 3 15 20 21 0.0166 0.0190

(20;25] 4 20 25 75 0.0592 0.0782

(25;30] 5 25 30 215 0.1698 0.2480

(30;35] 6 30 35 373 0.2946 0.5426

(35;40] 7 35 40 350 0.2765 0.8191

(40;45] 8 40 45 171 0.1351 0.9542

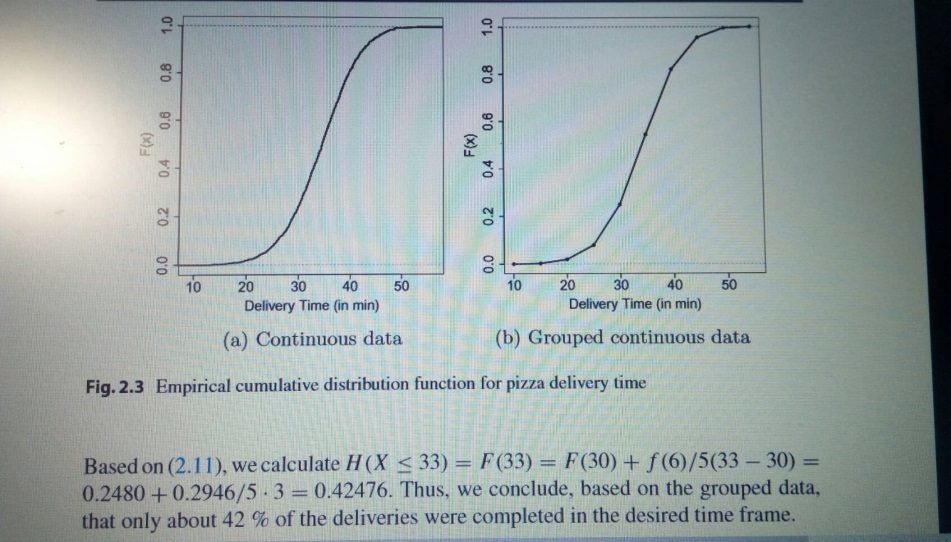
(45;50] 9 45 50 52 0.0411 0.9953

(50;55] 10 50 55 6 0.0047 1.0000

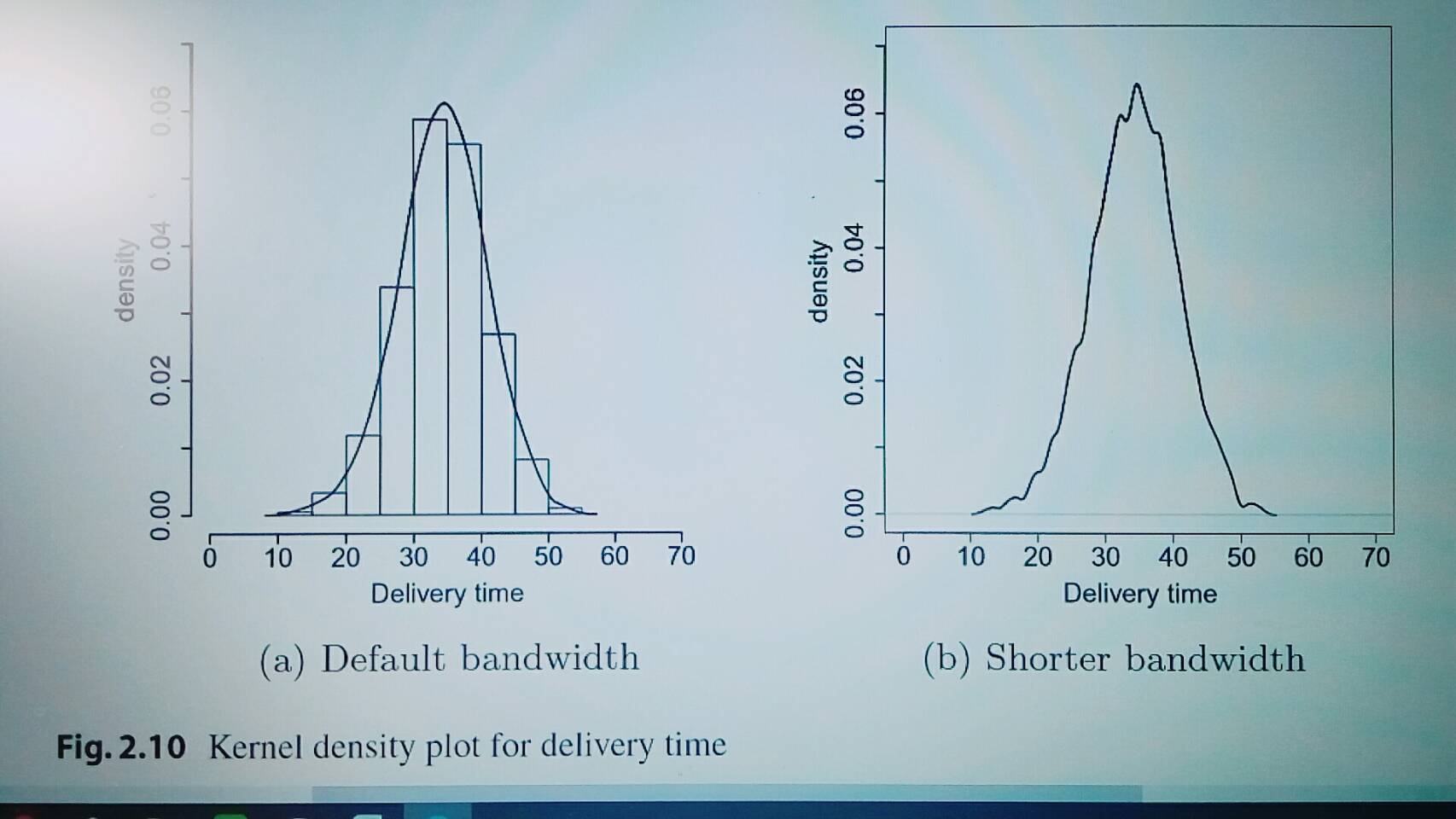
以下是上面那題的參考資料:

1. Example 2.2.3

Consider Example 2.1.3 of the pizza delivery service. Suppose we are interested in determining the distribution of the pizza delivery times. Using the function plot.ecdf() in R, we obtain the ECDF of the continuous data, see Fig.2.3a. Note that the structure of the curve is a step function but now almost looks like a continuous curve. The reason for this is that when the number of observations is large, then the lengths of class intervals becomes mall. When these small lengths are

1. Fig.2.3(下圖)

(還有這個圖要怎麼用R畫出來呢?)



這個圖要怎麼用R畫出來呢?

2.4

Load and attach the data set central .park (UsingR). The WX variable contains a list of numbers representing bad weather (e.g., 1 for fog, 3 for thunder, 8 for smoke or haze). NA is used when none of the types occurred.

Make a table(見2.) of the data, then

②make a table with the extra argument exclude=FALSE. Why is the second table better?

2.8

The data set npdb(這是什麼?是資料嗎?) (UsingR) contains information on malpractice awards in the United States. Attach the data set and make a table of the state variable. Which state had the most awards? (Using sort () on your table is useful here.)

下面這個連結是UsingR的電子書。我不懂題目要我做甚麼?

<https://ceiba.ntu.edu.tw/course/a79e21/bulletin/330750_2-Using_R_for_Introductory_Statistics.pdf>

2.9整個看不懂

For the malpractice-award data set npdb (UsingR), the variable ID is an identification number unique to a doctor but not traceable back to the doctor. It allows a look at a doctor’s malpractice record without sacrificing anonymity.

The commands

> table(npdb$ID)

create a table of malpractice awards for each of the 6,369 doctors. What does the command table (table (ID)) do, and why is this interesting?

2.16

The data set rivers contains the lengths (in miles) of 141 major rivers in North America.

1. What proportion are less than 500 miles long?

2. What proportion are less than the mean length?

3. What is the 0.75 quantile?

這整題的Data是我要去Using R裡面找嗎?

★題目說的Data要去哪裡找呢?Using的電子書上嗎?還是R程式裡面就有?★

2.17

The time variable in the nym. 2002 (UsingR) data set contains the time to finish the 2002 New York City marathon for a random sample of the finishers.

1. What percent ran the race in under 3 hours?

2. What is the time cutoff(是什麼意思?) for the top 10%? The top 25%?

3. What time cuts off the bottom 10%?

Do you expect this data set to be symmetrically distributed?

2.22

The median absolute deviation is defined as mad(x)=1.4826·median(|xi-median(x)|).

(2.5)

This is a resistant measure of spread and is implemented in the mad () function. Explain in words what it measures. Compare the values of the sample standard deviation, IQR, and median absolute deviation for the exec.pay (UsingR) data set.

2.25

We may prefer the standard deviation to measure spread over the variance as the units are the same as the mean. Some disciplines, such as ecology, prefer to have a unitless measurement of spread. The coefficient of variation is defined as the standard deviation divided by the mean.

One advantage is that the coefficient of variation matches our intuition of spread. For example, the numbers 1, 2, 3, 4 and 1001, 1002, 1003, 1004 have the same standard deviation but much different coefficient of variations. Somehow, we mentally think of the latter set of numbers as closer together.

For the rivers and pi2000 (UsingR) data sets, find the coefficient of variation.

2.44

It can be illuminating to view two different graphics of the same data set at once. A simple way to stack graphics is to specify that a figure will contain two graphics by using the command

> par(mfrow=c(2,1) # 2 rows, 1 column for graphic figures

Then, if x is the data set, the commands

> hist(x)

> boxplot(x, horizontal=TRUE)

will produce stacked graphics. (The graphics device will remain divided until you change it back with a command such as par (mfrow=c(1, 1)) or close the device.)

For the data set lawsuits (UsingR), make stacked graphics of lawsuits and log (lawsuits). Could you have guessed where the middle 50% of the data would have been without the help of the boxplot?

2.45

Sometimes a data set is so skewed that it can help if we transform the data prior to looking at it. A common transformation for long-tailed data sets is to take the logarithm of the data. For example, the exec.pay (UsingR) data set is highly skewed. Look at histograms before and after taking a logarithmic transform. Which is better at showing the data and why? (You can transform with the command log (1+exec. pay, 10).) Find the median and the mean for the transformed data. How do they correspond to the median and mean of the untransformed data?