GY7702_CW2

209040725

21/12/2020

This document was created to meet the requirements of GY7702 R for Data Science at University of Leicester. It was designed and created in R Markdown, a markup language that allows users to create documents that can be formatted to embed code blocks, code outputs and hyperlinks. When the R Markdown file is compiled, the markup language is hidden and the document is displayed in plain text.

This content was created using R, Rstudio, RMarkdown and GitHub

The libraries used in this assignment were

```
library(tidyverse)
library(pastecs)
library(knitr)
library(tinytex)
library(kableExtra)
```

Load in the data

```
# Extract
unzip("Data/GY7702_2020-21_Assignment_2--data_pack.zip", exdir = "Data")

# Load
OAC_2011 <-
  readr::read_csv(
    "Data/GY7702_2020-21_Assignment_2--data_pack/2011_OAC_Raw_kVariables.csv")

OAC_2011_meta <-
  readr::read_csv(
    "Data/GY7702_2020-21_Assignment_2--data_pack/OA11_LSOA11_MSOA11_LAD11_EW_LUv2.csv")
```

Option A

Joining the data sets

```
# Rename the column 'OA11CD' to OA so it matches the same column in OAC_2011
OAC_2011_meta <- OAC_2011_meta %>%
  dplyr::rename(
```

```
    OA = OA11CD
  )

# Join OAC_2011 with OAC_2011_meta, based on the OA column for each data frame
complete_OAC <-
  OAC_2011 %>%
  dplyr::left_join(.,OAC_2011_meta)
```

Subsetting

- Selecting assigned LAD
- Selecting the appropriate variables

```
# Subset data based off LAD11NM = Wellingborough
wellingborough <- complete_OAC %>%
  dplyr::filter(LAD11NM == "Wellingborough")

# Select the appropriate variables, based on their 'VariableCode'
wellingborough_variables <- wellingborough %>%
  dplyr::select(k004, k009, k010, k027, k031, k041, k046)
```

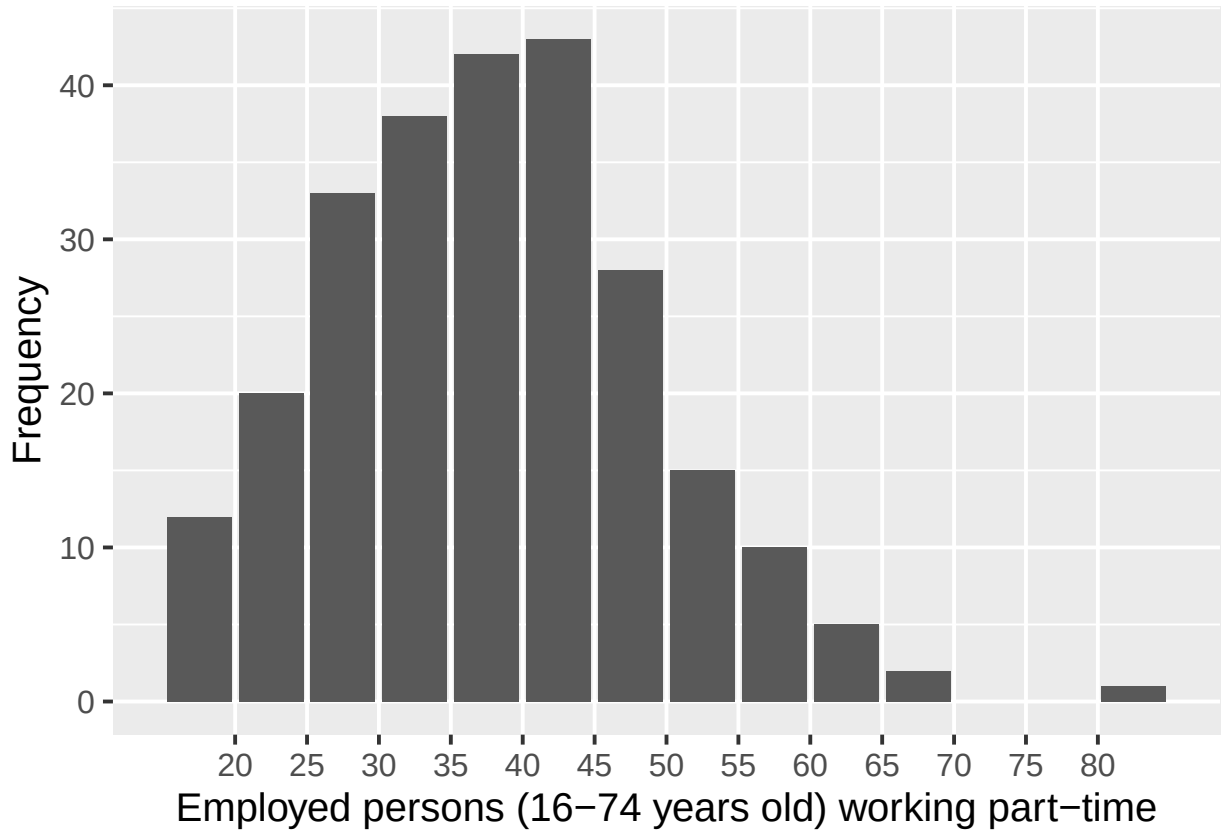
Question A.1

Exploring the distribution of Employed persons aged between 16 and 74 who work part-time

For this exploratory analysis, we are going to investigate the distribution of part-time employed persons (PTEP) throughout 249 output areas (OA's) in the Local Area District (LAD) of Wellingborough

Basic histogram for visual histogram

```
wellingborough %>%
  ggplot2::ggplot(aes(x = k046))+
  ggplot2::geom_bar()+
  scale_x_binned(n.breaks = 15)+
  xlab("Employed persons (16-74 years old) working part-time")+
  ylab("Frequency")+
  theme_gray(base_size = 15)
```



An initial assessment suggests that the distribution of PTEPs throughout Wellingborough is normal, that is bell-shaped. In other words, the majority of OA's in Wellingborough have 35-45 PTEPS. Some OA's have as little as 20 and at least one OA has >80 PTEPs.

We previously stated that most OA's have 35-45 PTEPs. Due to this, we would expect the average to be around this range. We can check this with

```
wellingborough_stats <- wellingborough %>%
  dplyr::select(k046)%>%
  pastecs::stat.desc(basic = FALSE, desc = TRUE, norm = FALSE)

kable(wellingborough_stats, format = "latex", booktabs = T) %>%
  kable_styling(latex_options = c("striped", "scale_down"))
```

	k046
median	38.0000000
mean	38.3293173
SE.mean	0.7237324
CI.mean.0.95	1.4254458
var	130.4233709
std.dev	11.4203052
coef.var	0.2979522

Note, this can also be achieved with

```
wellingborough%>%
  dplyr::summarise(Avg_PTEP = mean(k046))
```

Though the code can get lengthy with the addition of more and more statistics

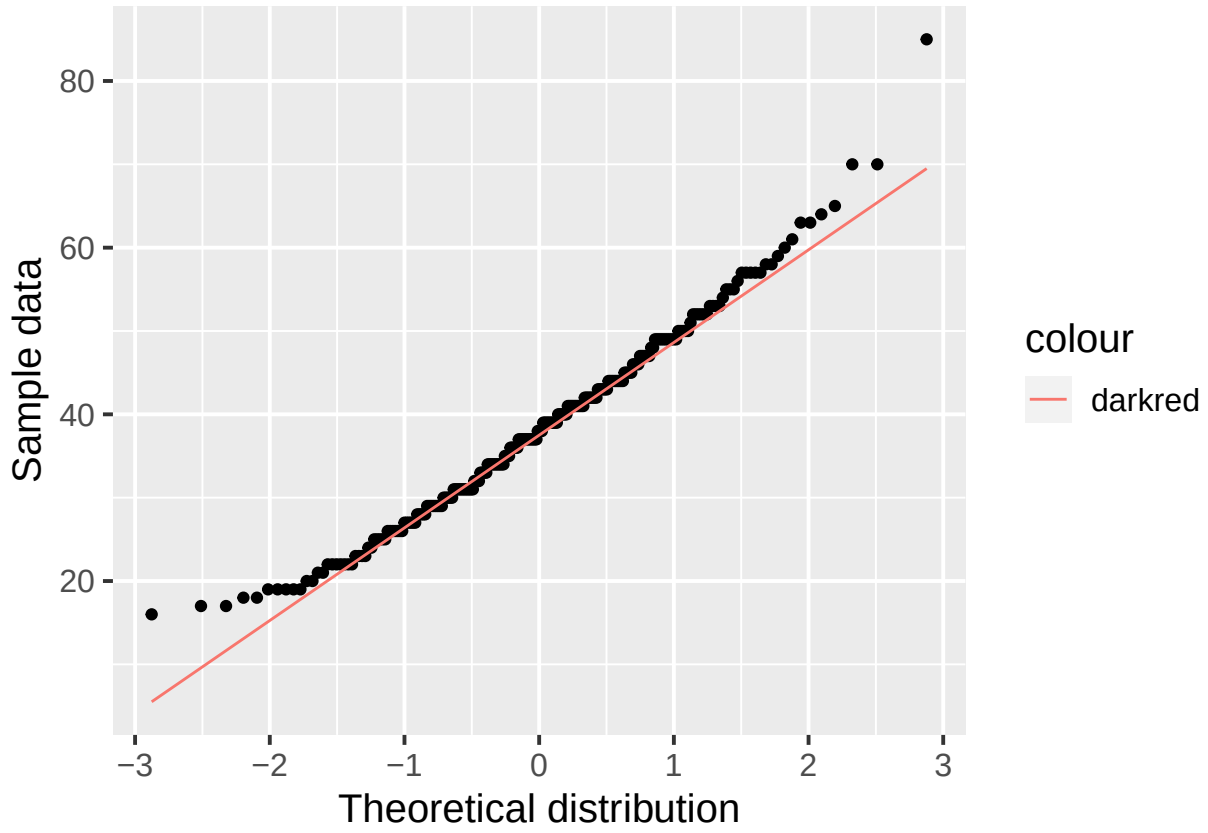
Plotting against normal distribution

We have seen that the distribution of PTEPs throughout Wellingborough looks normal, with an average (38 PTEPs per OA) tha sits neatly in the middle of the histogram.

The distribution can be assessed alternatively in a plot that gives insight into where values may lie that are affecting the normal distribution

```
wellingborough %>%
  ggplot2::ggplot(aes(sample = k046))+
  ggplot2::stat_qq()+
  ggplot2::stat_qq_line(
    aes(
```

```
     color = "darkred"))+
  xlab("Theoretical distribution")+
  ylab("Sample data")+
  theme_gray(base_size = 15)
```



The function `stat_qq` highlights a theoretical, normal distribution with the distribution of the PTEP data. This demonstrates that there are an elevated number of PTEPs at the tails of the distribution. That is, there are a considerable number of OA's with a minimum and maximum number of PTEPs. Additionally, notice the outlier on figure 1 here, >80.

Shapiro Wilko test

After visualising the data, we can quantitively assess if this is indeed a normal distribution or not

```
wellingborough %>%
  dplyr::pull(k046)%>%
  stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.98179, p-value = 0.002799
```

In the shapiro-wilk test, we set a null hypothesis that the data is normally distributed. The null hypothesis is rejected is the p-value <0.01. That is, there is a <1% of the data being normally distributed.In this case $p < 0.01$, the data is not normally distributed.

Exploring skewness and kurtosis

We can gain more information on the shape of the data by assessing its skew and kurtosis

```
wellingborough_skew <- wellingborough %>%
  dplyr::select(k046)%>%
  pastecs::stat.desc(basic = FALSE, desc = FALSE, norm = TRUE)

kable(wellingborough_skew, format = "latex", booktabs = T) %>%
kable_styling(latex_options = c("striped", "scale_up"))
```

	k046
skewness	0.4888972
skew.2SE	1.5841824
kurtosis	0.4670775
kurt.2SE	0.7596877
normtest.W	0.9817935
normtest.p	0.0027990

In this case, a positive sk ew means the data leans towards the left. The standard error for the skew `skew.2SE` also lies outside of the range -1 + 1, meaning the result is significant. A positive kurtosis also means the data has heavy tails. However, this result is not significant as the standard error `kurt.2SE` lies within -1 +1