


# Project

Candidate Number: 

## Profile of XYZ

The profile of XYZ:

- Age: 23
- Gender: Female
- Home address: Guildford

## Introduction and description of dataset

The goal of this project is to investigate the passing rates of the practical driving test at the various test driving test centres (DTCs) in the United Kingdom, since it is widely believed that driving test routes around some DTCs are probably more difficult to others. We wish to make a recommendation to XYZ as to which DTC should they take the test.

The dataset used for this analysis is the **DVSA1203**, downloaded from the ST447 Moodle page. This dataset is produced by the *Driver and Vehicles Standards Agency (DVSA)*. This dataset contains information on the number of tests conducted and the number of passes, reported by age (17 to 25 year olds), gender, year (between 2007/08 and 2021/22), and DTC.

The dataset is provided in the OpenDocument Spreadsheet (.ods) file format. The spreadsheet contains 16 separate sheets, with the first sheet being notes regarding this dataset, and the remaining 15 sheets contains data for each time period reported (2007/08 to 2021/22). We note that the data is not formatted in exactly the same way in each sheet. According to the notes given in the dataset, some of the information in the dataset is withheld and is indicated by “.” for privacy reasons and to prevent the identification of individuals. The data is also withheld if only one examiner has conducted that category of testing at a particular DTC. Nonetheless, the results of these tests are included in the aggregated results.

Using `readODS`, `tidyr` and `dplyr` packages, we import, then format and extract only the data relevant to the two DTCs we are interested in: Guildford (nearest to XYZ’s home), and Wood Green (nearest to LSE).

## Exploratory data analysis (EDA)

We first provide summary statistics regarding the passing rates of 23 year old females at the Guildford and Wood Green DTCs between 2007/08 and 2021/22. This can also be interpreted as XYZ’s expected passing rate at the two DTCs. From Table 1, the mean passing rate at the Guildford DTC is 0.4835, with a standard error of 0.0197; and the mean passing rate at Wood Green is 0.3828, with a standard error of 0.0141. The figures for the median passing rate are similar.

DTC	Median	Mean	Standard Error
Guildford	0.4925	0.4835	0.0197
Wood Green	0.3803	0.3828	0.0141

Table 1: Comparison of passing rate of 23 year old females at Guildford and Wood Green DTCs between 2007/08 and 2021/22.

Plots of the passing rates of 23 year old females at the Guildford and Wood Green DTCs are also shown below. Figure 1 plots the passing rate of 23 year females at the two DTCs between 2007/08 and 2021/22. We observe that the passing rate is higher at the Guildford DTC for 13 out of 15 years reported. Figure 2 shows the distribution of the passing rates between 2007/08 and 2021/22. We also note that the median and mean passing rates are higher at the Guildford DTC than the Wood Green DTC by 0.1.

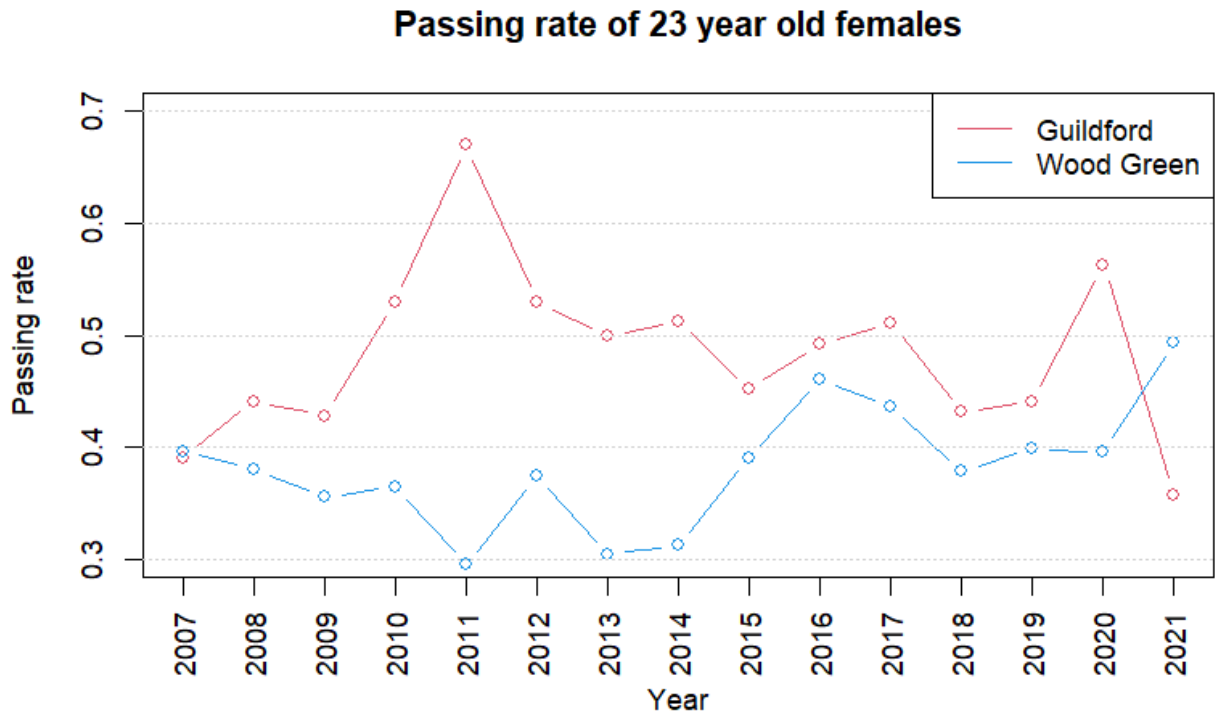


Figure 1: Line graph of passing rate of 23 year old females at Guildford and Wood Green DTCs between 2007/08 and 2021/22.

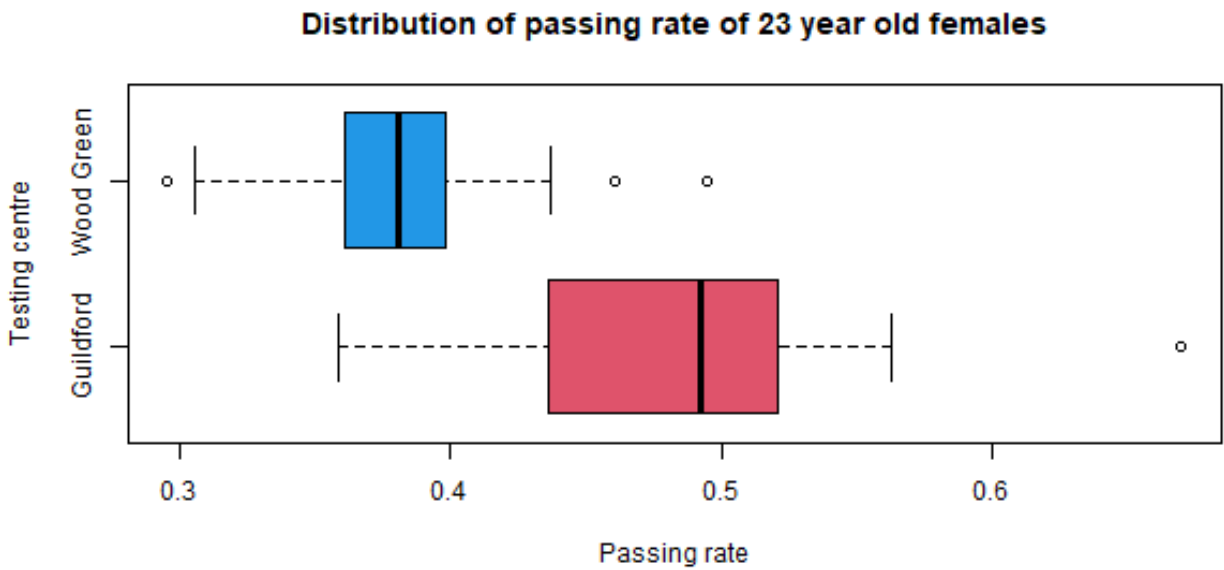


Figure 2: Box plot of the distribution of passing rates of 23 year old females at Guildford and Wood Green DTCs between 2007/08 and 2021/22.

## Methodology

Based on the EDA, it is reasonable to speculate that the passing rate at the Guildford DTC is higher than that of the Wood Green DTC. We will conduct a statistical hypothesis test defined as follows. We will make a number of assumptions for which the implications will be discussed in a later section.

Let  $G_1, \dots, G_n$  represent the observations (passing rate of 23 year old females for each year between 2007/08 and 2021/22) at the Guildford DTC. We will assume these observations are independent and identically distributed random variables sampled from a normal (Gaussian) distribution with mean  $\mu_G$  and variance  $\sigma_G^2$ . Similarly, let  $W_1, \dots, W_m$  represent the observations at the Wood Green DTC, sampled from a normal distribution with mean  $\mu_W$  and variance  $\sigma_W^2$ . We will assume that the two variances are equal ( $\sigma^2 = \sigma_G^2 = \sigma_W^2$ ). Therefore, the sample mean of the two samples also follow a normal distribution given as follows:

$$\left(\bar{G} = \frac{1}{n} \sum_{i=1}^n G_i\right) \sim N\left(\mu_G, \frac{\sigma^2}{n}\right) \quad \text{and} \quad \left(\bar{W} = \frac{1}{m} \sum_{i=1}^m W_i\right) \sim N\left(\mu_W, \frac{\sigma^2}{m}\right).$$

Thus, assuming the two samples are independent to each other, we have that:

$$\bar{G} - \bar{W} \sim N\left(\mu_G - \mu_W, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right) \implies \frac{(\bar{G} - \bar{W}) - (\mu_G - \mu_W)}{\sqrt{\sigma^2/n + \sigma^2/m}} \sim N(0, 1).$$

We will conduct a hypothesis test on the difference of the means of the two samples, i.e. whether the mean passing rates at the Guildford and Wood Green DTCs are significantly different. Let the null hypothesis  $H_0$  be  $\mu_G = \mu_W$ , and the alternative hypothesis  $H_1$  be  $\mu_G \neq \mu_W$ . Since the true population variance is unknown, we estimate it using the sample variance, defined as:

$$S_G^2 = \frac{1}{n-1} \sum_{i=1}^n (G_i - \bar{G})^2 \quad \text{and} \quad S_W^2 = \frac{1}{m-1} \sum_{i=1}^m (W_i - \bar{W})^2.$$

We note without proof, that the following quantities follow a  $\chi^2$ -distribution with  $n-1$  and  $m-1$  degrees of freedom, respectively. Therefore, their sum also follows a  $\chi^2$ -distribution with  $n+m-2$  degrees of freedom.

$$\frac{(n-1)S_G^2}{\sigma^2} \sim \chi_{n-1}^2 \quad \text{and} \quad \frac{(m-1)S_W^2}{\sigma^2} \sim \chi_{m-1}^2 \implies \frac{(n-1)S_G^2}{\sigma^2} + \frac{(m-1)S_W^2}{\sigma^2} \sim \chi_{n+m-2}^2. \quad (1)$$

The test statistic, which follows the (Student's)  $t$ -distribution with  $n+m-2$  degrees of freedom (written as  $T \sim t_{n+m-2}$ ), is therefore:

$$T = \frac{(\bar{G} - \bar{W}) - (\mu_G - \mu_W)}{\sqrt{\sigma^2/n + \sigma^2/m}} \bigg/ \sqrt{\frac{(n-1)S_G^2 + (m-1)S_W^2}{\sigma^2(n+m-2)}} = \sqrt{\frac{n+m-2}{1/n + 1/m}} \times \frac{(\bar{G} - \bar{W}) - (\mu_G - \mu_W)}{\sqrt{(n-1)S_G^2 + (m-1)S_W^2}} \sim t_{n+m-2}.$$

An (accurate)  $(1-\alpha)\%$  confidence interval for the difference of the means is given as follows:

$$\bar{G} - \bar{W} \pm t_{\alpha/2, n+m-2} \times \sqrt{\frac{1/n + 1/m}{n+m-2} \times ((n-1)S_G^2 + (m-1)S_W^2)}.$$

Therefore, under the null hypothesis  $H_0 : \mu_G = \mu_W$ , we calculate the  $t$ -value to be:

$$T = \sqrt{\frac{15+15-2}{1/15 + 1/15}} \times \frac{(0.4835 - 0.3828) - 0}{\sqrt{(15-1) \times 0.005834 + (15-1) \times 0.002997}} \approx 4.1470.$$

Using the function `qt()` in R (or from statistical tables), for  $T \sim t_{28}$ , we have that  $\mathbb{P}(T > 2.048407) = 0.025$ . Therefore, this test is significant at the 5% significance level, so we can reject the null hypothesis at the 5% significance level, i.e. the mean passing rates at Guildford and Wood Green DTCs are significantly different.

An (accurate) 95% confidence interval for the difference of the means is therefore:

$$(0.4835 - 0.3828) \pm 2.048 \times \sqrt{\frac{1/15 + 1/15}{15 + 15 - 2}} \times ((15 - 1) \times 0.005834 + (15 - 1) \times 0.002997) = (0.0509, 0.1503).$$

That is to say that, the difference of the means ( $\mu_W - \mu_G$ ) is estimated to be  $0.1006 \pm 0.0497$  with 95% confidence. This is consistent with our observations in the EDA, where the mean passing rate at the Guildford DTC is higher than that of the Wood Green DTC by about 0.1, which is contained in the confidence interval. The  $t$ -test to compare the difference of means can also be performed simply by using the function `t.test()` in R.

Therefore, I would recommend XYZ to take the driving test at the Guildford DTC.

## Model assumptions

In the analysis above, we have made a number of modelling assumptions, including normality and equality of variances. We now investigate the validity of these assumptions.

Firstly, regarding the assumption of normality, we can examine the normal quantile-quantile (Q-Q) plots, shown below in Figure 3. Most of the observations at the Guildford DTC appears to lie close to the straight line, where as the observations at the Wood Green DTC appear to deviate at both ends, but symmetrically. If we consider all the observations collectively, most appears to lie close to the straight line. Overall, we conclude that the data is well modelled by a normal distribution.



Figure 3: Normal quantile-quantile (Q-Q) plot of passing rates of 23 year old females between 2007/08 and 2021/22 at Guildford (left) and Wood Green (middle) DTCs, as well as collectively (right).

Secondly, regarding the assumption of equal variances, we can perform a  $F$ -test on the ratio of two variances. Again we assume that  $G$  and  $W$  follow a normal distribution. Using the result of (1), the test statistic which follows the  $F$ -distribution with  $n - 1$  and  $m - 1$  degrees of freedom (written as  $T \sim F_{n-1, m-1}$ ) is as follows:

$$T = \frac{S_G^2 / \sigma_W^2}{S_W^2 / \sigma_G^2} = \frac{\sigma_W^2}{\sigma_G^2} \times \frac{S_G^2}{S_W^2} \sim F_{n-1, m-1}.$$

The null hypothesis  $H_0$  is  $\sigma_G^2 / \sigma_W^2 = 1$ , and the alternative hypothesis  $H_1$  is  $\sigma_G^2 / \sigma_W^2 \neq 1$ . For the given data, under the null hypothesis  $H_0 : \sigma_G^2 / \sigma_W^2 = 1$ , the test statistic is calculated to be:

$$T = \frac{\sigma_W^2}{\sigma_G^2} \times \frac{S_G^2}{S_W^2} = 1 \times \frac{0.005834}{0.002997} = 1.946438.$$

Using the function `qf()` in R (or from statistical tables), for  $T \sim F_{14, 14}$ , we have that  $\mathbb{P}(T > 2.978588) = 0.025$ . Therefore, with the given data, we cannot reject  $H_0$ : we may not conclude that the variances are significantly different, at the 5% significance level. Hence, the assumption of equal variances is justified. The  $F$ -test to compare the ratio of two variances can also be performed simply by using the function `var.test()` in R.

## Strengths, weaknesses and limitations

In the previous section, we have performed a  $t$ -test instead of the Wald test. This has the advantage of the test statistic being more accurate when the number of observations ( $n$  and  $m$ ) are small. However, this does rely on the assumption that the underlying distribution of the passing rates is actually a normal distribution. The Wald test, on the other hand, only uses the fact that the test statistic asymptotically follows a (standard) normal distribution. Alternatively, we could have also performed the permutation test, which also does not rely on any asymptotic theory.

We have also assumed that (1) the test itself has remained unchanged (since 2007/08), for example, the types of manoeuvres assessed in the test; (2) the standards of the test, i.e. the level of performance or competency expected of the candidate; (3) the route of the driving test at each DTC has remain (largely) unchanged, for example, no significant changes to road layouts in the local area surrounding the DTC; and (4) that there have been no significant changes to traffic laws, which may have other implications.

The data for 2020/21 may also be considered as unrepresentative due to the COVID-19 pandemic, during which there were significantly fewer candidates as the driving tests were suspended during the lockdowns.

Further, we have only compared the data for 23 year old females. We can repeat the analysis for other ages and/or genders. If we observe a similar pattern, (that is, the passing rate at the Guildford DTC, is on average higher than that of the Wood Green DTC), then this can also be considered stronger evidence that XYZ should take their driving test at the Guildford DTC.

The data in the **DVSA1203** dataset does not distinguish the number of attempts a candidate has had at the driving test. The DVSA also reports the number of attempts at the driving test before passing in the **DRT0202** dataset, reported by age and gender but on a nation-wide (Great Britain) scale instead of by DTC. Elementary inspections of this dataset suggests that a candidate is more likely to pass their driving test in their first two attempts, with the rate of passing falling as they make more attempts at the test. Therefore, this may suggest that the data in **DVSA1203** dataset underestimates the passing rate for first-time candidates, assuming this is XYZ's first attempt at the driving test. The data in **DVSA1203** dataset also does not distinguish the ethnicity of the candidate. The **DVSA1204** dataset contains passing rates by ethnicity of candidate, per DTC. Although this should not impact the passing rate of candidates, elementary inspections of this dataset suggests otherwise.

Moreover, we have not considered where has XYZ taken their driving lessons. XYZ could be more familiar with the road layout and conditions in the area where they have taken driving lessons, and hence be more confident and possibly have better chance of passing the test, if they take the test at a DTC in the same area as where they took the lessons.

Finally, we have only considered the data from two DTCs. It is possible that there is some other DTC, for example, in between Guildford (nearest to XYZ's home) and Wood Green (nearest to LSE), where the passing rate there is more favourable. In such case, XYZ may prefer to take the test at an alternative DTC instead.

## R code used in this analysis

```
1 library(readODS) # for read_ods()
2 library(tidyr)   # for fill()
3 library(dplyr)   # for filter()
4
5 # Import data
6 path <- "dvsa1203.ods" # Assuming this file is placed in the same directory
7 sheets <- list_ods_sheets(path) # Get list of sheet names
8 sheets <- sheets[-1] #The first sheet is Notes
9 col_names <- c("Test_centre", "Age", "Male_count", "Male_pass", "Male_pass_rate", "
  Female_count", "Female_pass", "Female_pass_rate", "Total_count", "Total_pass", "
  Total_pass_rate")
10 data <- c()
11
12 for (sheet in sheets){
13   print(sheet) #Progress bar
14   dat <- c()
15   dat <- read_ods(path, sheet=sheet, col_names=TRUE, na="", skip=6) #Read sheet
16   empty_columns <- sapply(dat, function(x) all(is.na(x) | x=="")) #Find NA columns
17   dat <- dat[,!empty_columns] #Remove NA columns
18   colnames(dat) <- col_names #Rename columns
19   dat <- fill(dat, Test_centre, .direction="down") #Fill in test centre
20   dat <- filter(dat, Test_centre=="Guildford" | Test_centre=="Wood Green (London)" |
  Test_centre=="Wood Green") #Subset only the two relevant test centres
21   dat <- filter(dat, !is.na(Age), Age!="Total") #Remove empty rows and the total row
22   y <- unlist(strsplit(sheet, split="-", fixed=TRUE))[1]
23   yr <- rep(y, 18)
24   dat <- cbind(yr, dat) #Add column for years
25   data <- rbind(data, dat) #Merge data frames
26 }
27
28 # Convert data type to numeric
29 data[,c(1,3:12)] <- sapply(data[,c(1,3:12)], function(x) as.numeric(x))
30
31 # Make names of test centres consistent
32 data$Test_centre[data$Test_centre == "Wood Green"] <- "Wood Green (London)"
33
34 # Convert percentages to decimal
35 data$Male_pass_rate <- data$Male_pass_rate / 100
36 data$Female_pass_rate <- data$Female_pass_rate / 100
37 data$Total_pass_rate <- data$Total_pass_rate / 100
38
39 # Subset dataframe by relevant test centre, age and gender
40 dfg <- filter(data, Test_centre=="Guildford", Age==23)
41 dfw <- filter(data, Test_centre=="Wood Green (London)", Age==23)
42 dfg_1 <- dfg[,c("yr", "Female_pass_rate")]
43 dfw_1 <- dfw[,c("yr", "Female_pass_rate")]
44
45 # Rename columns and merge the two dataframes
46 colnames(dfg_1)[colnames(dfg_1)=="Female_pass_rate"] <- "Guildford"
47 colnames(dfw_1)[colnames(dfw_1)=="Female_pass_rate"] <- "Wood Green (London)"
48 dff <- merge(dfg_1, dfw_1)
49
50 # Line plot
51 png(filename="line.png", width=800, height=400, bg="white")
52 par(cex=1.5)
53 xlim <- c(2007, 2021)
54 ylim <- c(0.3, 0.7)
55 plot(dff$yr, dff$Guildford, "b", col=2, xlim=xlim, ylim=ylim, xlab="Year", ylab="
  Passing rate", xaxt="n", main="Passing rate of 23 year old females")
56 lines(dff$yr, dff$'Wood Green (London)', "b", col=4, xlim=xlim, ylim=ylim)
57 grid(nx=NA, ny=NULL, col="lightgray", lty="dotted", lwd=0.5)
58 legend("topright", legend=c("Guildford", "Wood Green"), col=c(2,4), lty=1)
59 axis(1, at = seq(2007, 2022, by=1), las=2)
60
```

```

61 # Box plot
62 png(filename="box.png", width=800, height=400, bg="white")
63 par(cex=1.5)
64 boxplot(dfg_1$Guildford, dfw_1$'Wood Green (London)', horizontal=TRUE, xlab="Passing
    rate", ylab="Testing centre", main="Distribution of passing rate of 23 year old
    females", col=c(2,4), names=c("Guildford", "Wood Green"))
65
66 # Summary statistics
67 standard_error <- function(x){sd(x)/sqrt(length(x))}
68 median_g <- median(dfg_1$Guildford) # 0.4925373
69 median_w <- median(dfw_1$'Wood Green (London)') # 0.3803419
70 mean_g <- mean(dfg_1$Guildford) # 0.4834553
71 mean_w <- mean(dfw_1$'Wood Green (London)') # 0.3828296
72 var_g <- var(dfg_1$Guildford) # 0.005834305
73 var_w <- var(dfw_1$'Wood Green (London)') # 0.002997427
74 se_g <- standard_error(dfg_1$Guildford) # 0.01972191
75 se_w <- standard_error(dfw_1$'Wood Green (London)') # 0.01413607
76
77 # t-test
78 sqrt(28/((2/15)))*((mean_g-mean_w)/(sqrt(14*(var_g+var_w)))) # 4.146977
79 ts <- qt(0.025, 28, lower.tail=FALSE) # t-value: 2.048407
80 # Confidence interval
81 (mean_g-mean_w)+ts*sqrt((2/15)/28*(14*(var_g+var_w))) # 0.15033
82 (mean_g-mean_w)-ts*sqrt((2/15)/28*(14*(var_g+var_w))) # 0.05092145
83 # Alternatively, we can use the t.test() function to obtain the same results
84 t.test(dfg_1$Guildford, dfw_1$'Wood Green (London)', var.equal=T)
85 #####
86 ## Two Sample t-test
87 ##
88 ## data: dfg_1$Guildford and dfw_1$'Wood Green (London)'
89 ## t = 4.147, df = 28, p-value = 0.000283
90 ## alternative hypothesis: true difference in means is not equal to 0
91 ## 95 percent confidence interval:
92 ## 0.05092144 0.15032995
93 ## sample estimates:
94 ## mean of x mean of y
95 ## 0.4834553 0.3828296
96 #####
97
98 # F-test
99 var_g / var_w # 1.946438
100 qf(0.025, 14, 14, lower.tail=FALSE) # F-value: 2.978588
101 # Alternatively, we can use the var.test() function to obtain the same results
102 var.test(dfg_1$Guildford, dfw_1$'Wood Green (London)')
103 #####
104 ## F test to compare two variances
105 ##
106 ## data: dfg_1$Guildford and dfw_1$'Wood Green (London)'
107 ## F = 1.9464, num df = 14, denom df = 14, p-value = 0.2251
108 ## alternative hypothesis: true ratio of variances is not equal to 1
109 ## 95 percent confidence interval:
110 ## 0.6534768 5.7976356
111 ## sample estimates:
112 ## ratio of variances
113 ## 1.946438
114 #####
115
116 ## QQ-plot
117 png(filename="qq.png", width=1200, height=400, bg="white")
118 par(mfrow=c(1,3), cex=1.5)
119 qqnorm(dfg_1$Guildford, main="Normal Q-Q Plot \n (Guildford)")
120 qqline(dfg_1$Guildford, col=2)
121 qqnorm(dfw_1$'Wood Green (London)', main="Normal Q-Q Plot \n (Wood Green)")
122 qqline(dfw_1$'Wood Green (London)', col=4)
123 qqnorm(c(dfg_1$Guildford, dfw_1$'Wood Green (London)'))
124 qqline(c(dfg_1$Guildford, dfw_1$'Wood Green (London)'))

```