7CCMMS61- Statistics for Data Analysis

Yongxin Cai

November, 2022

Question 1

1.a Crime Data Summary

	CrimeRate	CrimeRate10	ExpenditureYear0	ExpenditureYear10	Wage	Wage10
Minimum	45.50	26.50	45.00	41.00	288.00	359.00
Maximum	161.80	178.20	166.00	157.00	689.00	748.00
1. Quartile	82.70	76.35	62.50	58.50	459.50	530.00
3. Quartile	120.65	130.25	104.50	97.00	591.50	659.50
IQR	37.95	53.90	42.00	38.50	132.00	129.50
Mean	102.81	102.07	85.00	80.23	525.38	594.64
Median	103.00	103.50	78.00	73.00	537.00	615.00
Stdev	28.89	39.28	29.72	27.96	96.49	93.75
Skewness	0.13	0.01	0.89	0.84	-0.38	-0.43
Kurtosis	-0.71	-0.81	0.16	0.01	-0.61	-0.46

Table 1: Numerical summary table for the continuous variables

There are six continuous variables in the crime data. The following summary is based on the statistical (*Table 1*) and graphical representations (*Figure 1*) of the data.

- **CrimeRate** has values between 45.50 and 161.80. The middle 50% of CrimeRate values has a range 37.95, lie on (82.70, 120.65). Its standard deviation is 28.89, and there is no outlier shown by the boxplot (*Figure 1 top left*), so the CrimeRate data has a wide spread. It has an almost equal mean (102.81) and median (103.00). Together with this, the boxplot and probability density plot (*Figure 1 top right*) show that this variable has a nearly symmetrical distribution with a skewness of only 0.13. The Kurtosis value (-0.71) indicates that both the peak and the tail of CrimeRate are short. Based on these observations, it is likely to be a platykurtic distribution.[1]
- **CrimeRate10** has almost same values of mean and median as CrimeRate. But it has more wide range (26.50, 178.20) compared to CrimeRate. The middle 50% of CrimeRate10 has a range 53.90. It also has a large value of standard deviation (39.28). These numerical summaries implies CrimeRate10 is more spread out than CrimeRate. The boxplot and probability density plot (*Figure 1 top left and right*) also represent this. It has a longer box (i.e. wide range) and a flatter probability density plot (i.e. large sd). Based on its skewness and Kurtosis, it is symmetrical and light-tailed. It is likely to be a platykurtic distribution.[1]
- **ExpenditureYear0** has minimum 45.00 and maximum 166.00. The middle 50% values has a range 42.00. Together with its standard deviation 29.72, it has a greater deviation. Its mean (85.00) is greater than median (78.00), which indicates a skewness. It has a high positive value of skewness 0.89, a right skew and a right heavy tail can be seen in both boxplot and probability density plot (*Figure 1 middle left and right*). This means it is more likely to have a low value of expenditure on police in the 47 US states.
- **ExpenditureYear10** has lower overall values compared to ExpenditureYear0, which can be observed directly from its boxplot (*Figure 1 middle left*). There is an extreme value (i.e. the maximum 157.00) in ExpenditureYear10. By comparing its plot and table with ExpenditureYear0, this distribution is the same as ExpenditureYear0, i.e. right skew, a heavy tail on the right, wide spread.
- **Wage** has a range between 288.00 and 689.00. The middle 50% values are from 459.50 to 591.50, and it has a high value of standard deviation 96.49. A wide dispersion can be seen from its probability density plot (*Figure 1 bottom right*). Its mean 525.38 is less than its median 537.00 which indicates an asymmetric distribution. It has a moderate negative skewness (-0.38), a left skew and a left heavy tail can be seen in both boxplot (*Figure 1 bottom left*) and probability density plot, indicating that it is more likely to have a higher value for the median weekly wage across the 47 states in the US.
- **Wage10** has higher overall values compared to Wage, which implies an increase in wages over the decade. It can be observed directly from its boxplot (*Figure 1 bottom left*), Wage10 has a similar shape

of box but shifts upwards. It also has a same shape of probability density plot as Wage (*Figure 1 bottom right*), but shifts to the right. It can be concluded that Wage10 has a same ditribution as Wage, i.e. left shew, left heavy tailed, wide spread.

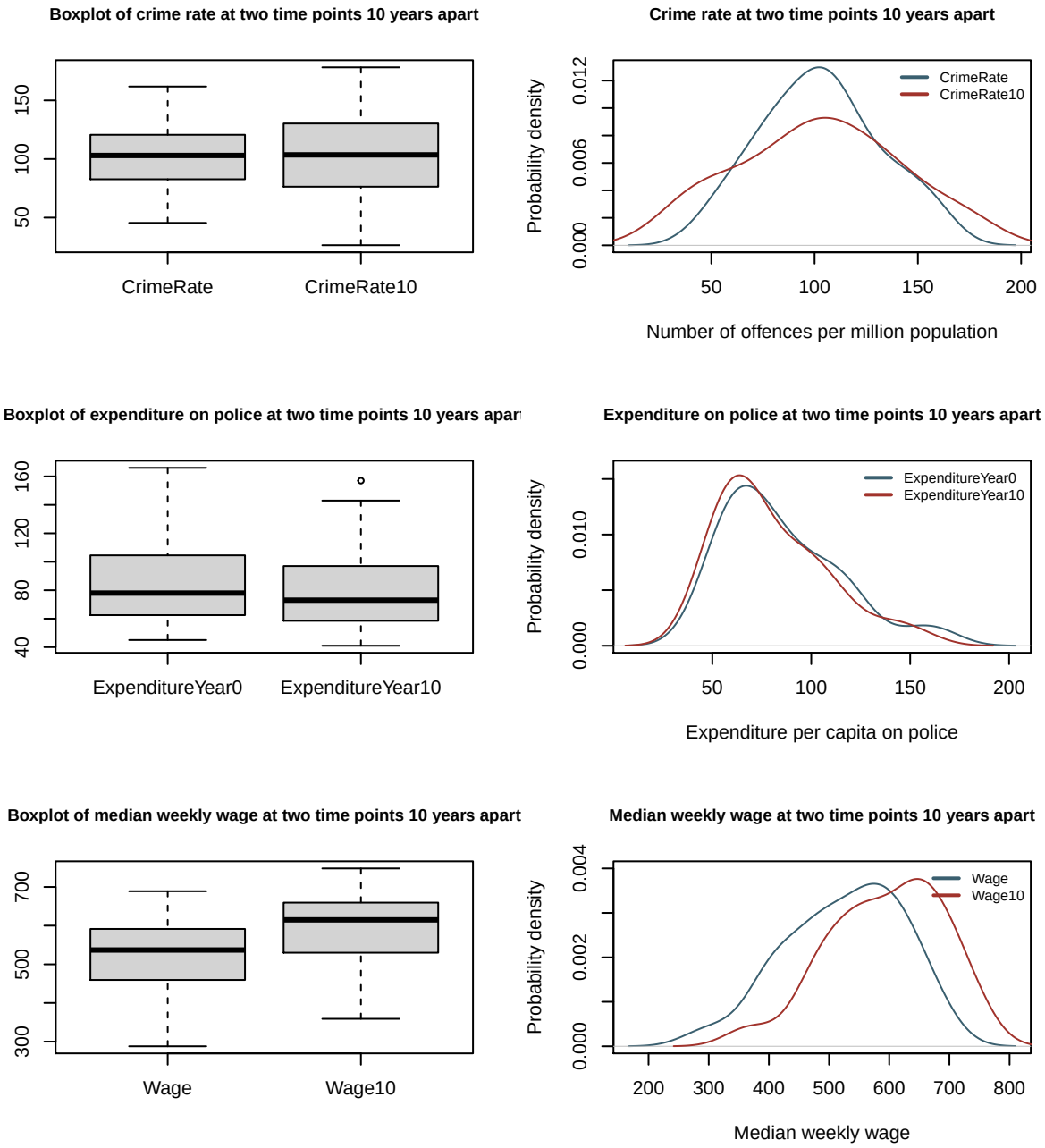

Figure 1: Boxplots and probability density plots of continuous variables in crime data

1.b Crime rate change over the decade

There is a change in crime rate over the decade. Based on the distribution in *Figure 1 top right*, it can be seen that the probability density plot has a lower peak and a heavier tail on both side compared to 10 years before, it has a greater deviation. The result why this happened can be deduced by looking at *Figure 2*. The states with crime rate below 100 becomes less criminal after ten years. Conversely, states with crime rates

above 110 become more criminal after ten years. This makes the crime rate data still have the same mean and median but more spread out.

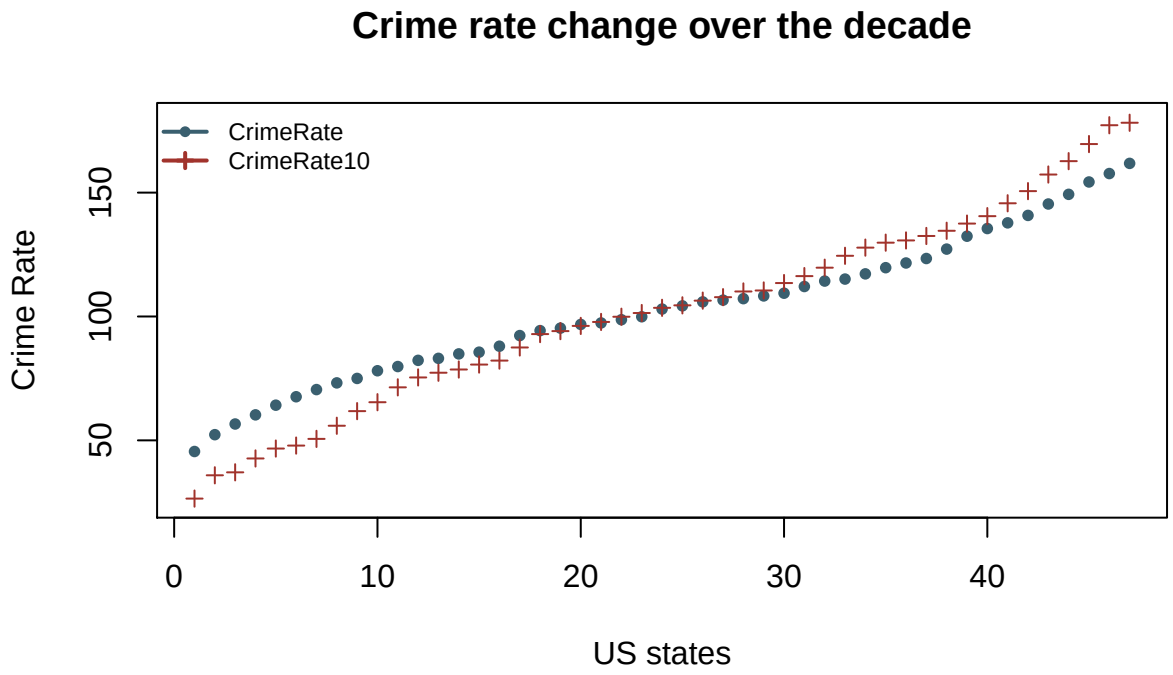

Figure 2: Changes in crime rates in each state over the decade

The correlation coefficient (*Table 2*) yields that there are a number of factors that affect crime rates. It can be showed that crime rate has a strong positive correlation with expenditure on police, and a moderate positive correlation with wage, and a weak negative correlation with high youth unemployment, and a weak positive correlation with state size. Since the last two variables have weak effects, and only 4/47 states have a decrease in state size, only 6/47 states have changes on factor level of HighYouthUnemploy, similarly, since there is a increase in wage in all 47 states over the decade and it only has a moderate correlation, we are not going to consider the influence from these three variables, we are only going to consider the effect from expenditure on police.

	ExpenditureYear0	StateSize	HighYouthUnemploy	Wage
CrimeRate	0.646	0.308	-0.286	0.425
	ExpenditureYear10	StateSize10	HighYouthUnemploy10	Wage10
CrimeRate10	0.63	0.304	-0.281	0.437

Table 2: Correlation coefficient of crime rate

The correlation coefficient shows that an increase in police expenditure is accompanied by an increase in crime rates. States with high expenditure on police is more likely to have high crime rates. Overall police expenditure has decreased by 6.4% compared to ten years ago, which has resulted in a slight decrease (0.74) in the average crime rate.
In summary, crime rates have become more dispersed, and it is likely that the reduced police expenditure per capita has led to a slight reduction in the overall crime rate.

Question 2

2.a

2.a.i Sketch the two species pine trees

The two distributions of pine trees are shown in *Figure 3*.

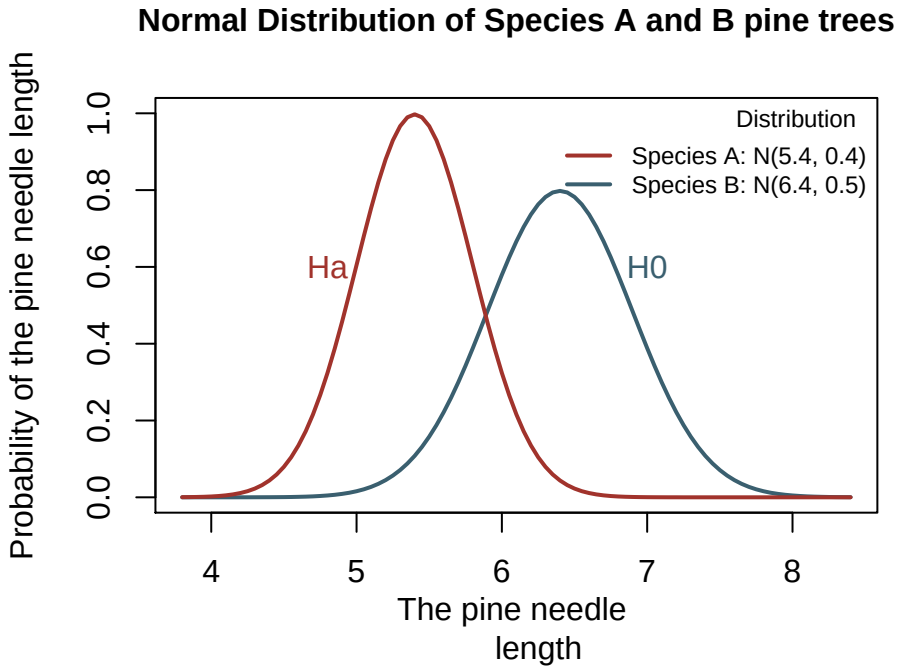

Figure 3: Normal distribution of species A and B pine trees

2.a.ii Direction of the 'extreme'

The direction of 'extreme' is one-sided to the left. The mean, μ_α, under H_α is assumed to be 5.4 which is lower than 6.4cm.

2.a.iii p-value

p-value = probability of the pine needle measures 5.42cm or less under H_0.

$$p - value = P(X \leq 5.42) = P(\frac{X - \mu_0}{\sigma} \leq \frac{5.42 - 6.4}{0.5}) = P(Z \leq -1.96) = 0.025$$

2.a.iv Conclusion

We have α=0.05 and p-value=0.025, the p-value is less than the level of significance α, so we reject H_0.
In conclusion, based on the statistical evidence from the sample, it can be shown statistically that the pine needle found in the suspect's coat would be very unlikely to find from Species B pine tree, thus, I claim that the pine needle found on the suspect is from Species A pine tree. The suspect was lying, he had been hunting in North Forest on the day of the incident.

Question 2

2.b

2.b.i State and sketch the two hypothesis

Let X be the demand for tickets for concerts at the O_2 arena in London, let $\overline{X}$ be the sample mean of X. The parameter of interest is the true mean of X, i.e.μ.
We have the following hypothesis:

$$H_0 : \mu = \mu_0$$

$$vs$$

$$H_\alpha : \mu > \mu_0$$

where $\mu_0 = 37$.
We have the known standard deviation $\sigma = 21$, since the sample size $n = 49$, we have $H_0 : \overline{X} \sim N(37, 21/\sqrt{49})$ which is $H_0 : \overline{X} \sim N(37, 3)$. The two hypothesis is shown at the top left of *Figure 4*.

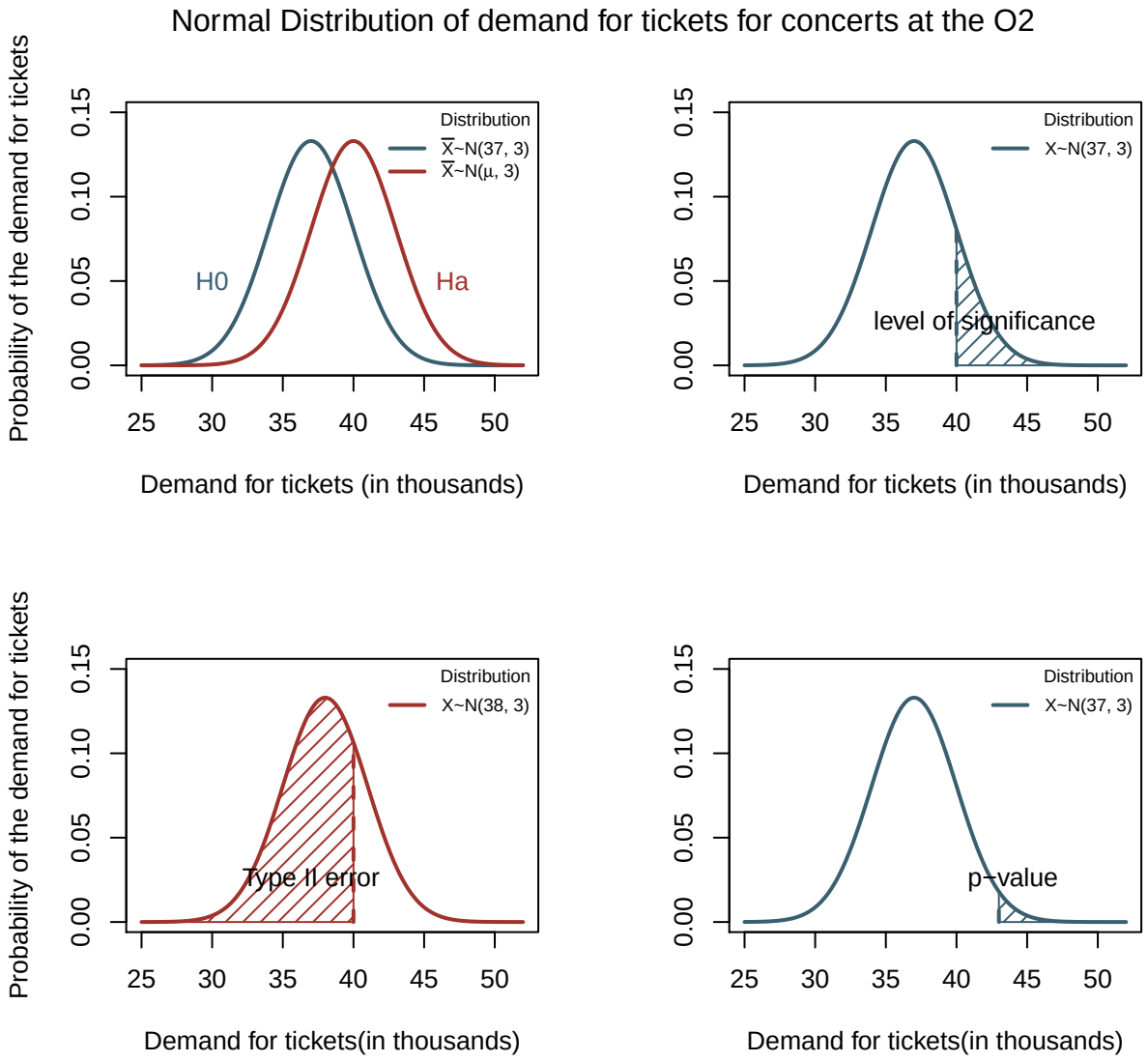

Figure 4: Normal distribution of demand for tickets for concerts at the O2 (top left), level of significance (top right), type II error (bottom left), p-value (bottom right)

2.b.ii Direction of the 'extreme'

The O_2 arena manager thought that mean demand(in thousands) for tickets for concerts, 37, is too low. He believes this value has increased, and is greater than 37. So the direction of the 'extreme' is one-sided to the right.

2.b.iii The level of significance

The decision rule, set up by the O_2 arena manager, he is going to proceed with the refurbishment if average ticket demand is greater than 40. So we have level of significance:

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ True}) = P(X > 40) = P(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > \frac{40 - 37}{\frac{21}{\sqrt{49}}}) = P(Z > 1) = 1 - P(Z < 1) = 0.159$$

The level of significance is shown at the top right of *Figure 4*.

2.b.iv Type II error

The probability of type II error:

$$\beta = P(\text{Reject } H_\alpha \mid H_\alpha \text{ True}) = P(X < 40) = P(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < \frac{40 - 38}{\frac{21}{\sqrt{49}}}) = P(Z < \frac{2}{3}) = 0.748$$

The type II error is shown at the bottom left of *Figure 4*.

2.b.v

Let Y be the number of null hypothesis been rejected, the probability of reject H_0 is $\alpha = 0.159$. Since there are 3 repeated and independent experiments, we have $Y \sim B(3, 0.159)$. The probability of at least two rejection is

$$P(Y \geq 2) = P(Y = 2) + P(Y = 3) = \binom{3}{2} 0.159^2 (1 - 0.159)^1 + \binom{3}{3} 0.159^3 (1 - 0.159)^0 = 0.068$$

2.b.vi p-value

p-value = probability of the ticket demand to be 43 or more under H_0.

$$p - value = P(X \geq 43) = P(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq \frac{43 - 37}{\frac{21}{\sqrt{49}}}) = P(Z \geq 2) = 1 - P(Z < 2) = 0.023$$

The p-value is shown at the bottom right of *Figure 4*.

2.b.vii Conclusion

We have α=0.252 and p-value=0.023, the p-value is less than the level of significance α, so we reject H_0. In conclusion, based on the evidence the O_2 arena manager has, it can be shown statistically that the average ticket is unlikely to have the original mean value, thus, it can be claimed that the average demand is greater than 40 and he should refurbish the arena.

Reference

[1] 'Kurtosis Definition, Types, and Importance', Investopedia [Online] Avaliable: https://www.investopedia .com/terms/k/kurtosis.asp

Appendix

Code

```
knitr::opts_chunk$set(echo = TRUE)
data_crime <- read.csv("crime.csv")
library(fBasics)
library(knitr)
library(formattable)
library(kableExtra)
attach(data_crime)
# Table 1
# Numerical summary table for the continuous variables
a = basicStats(CrimeRate)
b = basicStats(ExpenditureYear0)
c = basicStats(Wage)
d = basicStats(CrimeRate10)
e = basicStats(ExpenditureYear10)
f = basicStats(Wage10)
stat_summary = cbind(a, d, b, e, c, f)[c(3, 4, 5, 6, 7, 8, 14, 15, 16),]
stat_summary[10,]=stat_summary[4,] - stat_summary[3,]
rownames(stat_summary)[10] = 'IQR'
stat_summary = stat_summary[c(1,2,3,4,10,5,6,7,8,9),]
stat_summary %>% kable(caption = 'Numerical summary table for the continuous variables',
                       digits = 2) %>% kable_styling(position =
                       "center", latex_options = "hold_position")

# Figure 1
# Boxplots and probability density plots of continuous variables in crime data
par(mfrow=c(3,2))


# crime rate
boxplot(CrimeRate, CrimeRate10, names = c('CrimeRate', 'CrimeRate10'), main =
          'Boxplot of crime rate at two time points 10 years apart',
        cex.main = 0.9)

plot(density.default(x = CrimeRate, na.rm = TRUE), col = "#3a5f6f", main =
       'Crime rate at two time points 10 years apart',
       xlab = 'Number of offences per million population', ylab = 'Probability density',
     cex.main = 0.9)
lines(density.default(x = CrimeRate10, na.rm = TRUE), col = '#a1332c')
# Adding a legend
legend("topright", legend = c("CrimeRate", "CrimeRate10"), col = c("#3a5f6f",
                                                                   "#a1332c"),
       cex = 0.75,
       title.adj = 0.9, lty = 1, lwd = 2, box.lty = 0)


# expenditure
boxplot(ExpenditureYear0, ExpenditureYear10, names = c('ExpenditureYear0',
                                                       'ExpenditureYear10'),
        main = 'Boxplot of expenditure on police at two time points 10 years apart',
        cex.main = 0.9)
```

```r
plot(density.default(x = ExpenditureYear0, na.rm = TRUE), col = "#3a5f6f",
     main = 'Expenditure on police at two time points 10 years apart', xlab =
       'Expenditure per capita on police', ylab = 'Probability density', ylim =
       c(0, 0.016), cex.main = 0.9)
lines(density.default(x = ExpenditureYear10, na.rm = TRUE), col = '#a1332c')
# Adding a legend
legend("topright", legend = c("ExpenditureYear0", "ExpenditureYear10"), col =
         c("#3a5f6f", "#a1332c"), cex = 0.75,
       title.adj = 0.9, lty = 1, lwd = 2, box.lty = 0)


# wage
boxplot(Wage, Wage10, names = c('Wage', 'Wage10'),
    main = 'Boxplot of median weekly wage at two time points 10 years apart',
    cex.main = 0.9)

plot(density.default(x = Wage, na.rm = TRUE), col = "#3a5f6f",
    main = 'Median weekly wage at two time points 10 years apart',
    xlab = 'Median weekly wage', ylab = 'Probability density', ylim = c(0, 0.004),
    cex.main = 0.9)
lines(density.default(x = Wage10, na.rm = TRUE), col = '#a1332c')
# Adding a legend
legend("topright", legend = c("Wage", "Wage10"), col = c("#3a5f6f", "#a1332c"),
       cex = 0.75,
       title.adj = 0.9, lty = 1, lwd = 2, box.lty = 0)

# Figure 2
# Changes in crime rates in each state over the decade

c_r = as.data.frame(cbind(c(1:47), CrimeRate, CrimeRate10))
colnames(c_r)[1] = 'States'
plot(c_r$States, c_r$CrimeRate, xlab = 'US states', ylab = 'Crime Rate', ylim =
       c(25, 180), col = '#3a5f6f', pch=16, cex=0.8,
     main='Crime rate change over the decade')
points(c_r$States, c_r$CrimeRate10, col='#a1332c', pch =3, cex=0.8)
legend("topleft", legend = c("CrimeRate", "CrimeRate10"), col = c("#3a5f6f",
                                                                  "#a1332c"),
       cex = 0.75,
       title.adj = 0.9, pch = c(16, 3), lwd = 2, box.lty = 0)
# Table 2
# Correlation coefficient of crime rate

ceoff = cor(data_crime)[c(1,15),-c(1,15)][, c(4,8,11,12,16,20,23,24)]
ceoff = round(ceoff,3)
ceoff_table = rbind(ceoff[1,c(1:4)], colnames(ceoff)[5:8], ceoff[1,c(5:8)])
colnames(ceoff_table) = colnames(ceoff)[1:4]
rownames(ceoff_table) = c('CrimeRate', '', 'CrimeRate10')
ceoff_table %>% kable(caption = 'Correlation coefficient of crime rate') %>%
  kable_styling(position = "center", latex_options = "hold_position")
# Figure 3
# Normal distribution of species A and B pine trees

# Grid of X-axis values
```

```r
x <- seq(3.8, 8.4, 0.05)

# Plot
# Mean 6.4, sd 0.5
plot(x, dnorm(x, mean = 6.4, sd = 0.5), type = "l", xlab = 'The pine needle
     length', ylab = "Probability of the pine needle length", ylim = c(0, 1),
     lwd = 2, col = "#3a5f6f",
     main="Normal Distribution of Species A and B pine trees", cex.main=1)
# Mean 5.4, sd 0.4
lines(x, dnorm(x, mean = 5.4, sd = 0.4), col = "#a1332c", lty = 1, lwd = 2)
text(x=4.8, y=0.6, 'Ha', col="#a1332c")
text(x=7.0, y=0.6, 'H0', col="#3a5f6f")
# Adding a legend
legend("topright", legend = c("Species A: N(5.4, 0.4)", "Species B: N(6.4, 0.5)"
                              ), col = c("#a1332c", "#3a5f6f"),
       title = 'Distribution', cex = 0.75,
       title.adj = 0.9, lty = 1, lwd = 2, box.lty = 0)
# Figure 4
# Normal distribution of demand for tickets for concerts at the O2 (top left),
# level of significance (top right), type II error (bottom left), p-value (bottom right)

# Grid of X-axis values
x <- seq(25, 52, 0.1)

par(mfrow=c(2,2))

# Normal distribution of demand for tickets for concerts at the O2
# Mean 37, sd 3
plot(x, dnorm(x, mean = 37, sd = 3), type = "l", xlab = 'Demand for tickets (in thousands)',
     ylab = "Probability of the demand for tickets", ylim = c(0,
     0.15), lwd = 2, col = "#3a5f6f")
# Mean 40, sd 3
lines(x, dnorm(x, mean = 40, sd = 3), col = "#a1332c", lty = 1, lwd = 2)
text(x=30, y=0.05, 'H0', col="#3a5f6f")
text(x=47, y=0.05, 'Ha', col="#a1332c")
# Adding a legend
legend("topright", legend = c(expression(bar(X)*'~N(37, 3)'),
                              expression(bar(X)*'~N(' * mu * ', 3)')),
       col = c("#3a5f6f", "#a1332c"),
       title = 'Distribution', cex = 0.75,
       title.adj = 0.9, lty = 1, lwd = 2, box.lty = 0)


# Level of significance
# Mean 37, sd 3
plot(x, dnorm(x, mean = 37, sd = 3), type = "l",
     xlab = 'Demand for tickets (in thousands)', ylab = '', ylim = c(0, 0.15),
     lwd = 2, col = "#3a5f6f")
# Level of significance
segments(40, 0, 40, dnorm(40, 37, 3), lty = 2, lwd = 2, col = '#3a5f6f')
shade_x = c(seq(40,52,0.1), seq(52,40,-0.1))
shade_y = c(dnorm(seq(40,52,0.1), 37, 3), rep(0, length(shade_x)/2))
polygon(x = shade_x, y = shade_y, col = "#3a5f6f", density = 15, angle = 45)
```

```
text(x=42, y=0.025, 'level of significance')
# Adding a legend
legend("topright", legend = "X~N(37, 3)", col = "#3a5f6f",
       title = 'Distribution', cex = 0.75,
       title.adj = 0.9, lty = 1, lwd = 2, box.lty = 0)


# Type II error
# Mean 38, sd 3
plot(x, dnorm(x, mean = 38, sd = 3), type = "l", xlab = 'Demand for tickets(in thousands)',
     ylab = "Probability of the demand for tickets", ylim = c(0,
     0.15), col = "#a1332c", lty = 1, lwd = 2)
# Type II error
segments(40, 0, 40, dnorm(40, 38, 3), lty = 2, lwd = 2, col = '#a1332c')
shade_x = c(seq(25,40,0.1), seq(40,25,-0.1))
shade_y = c(dnorm(seq(25,40,0.1), 38, 3), rep(0, length(shade_x)/2))
polygon(x = shade_x, y = shade_y, col = "#a1332c", density = 15, angle = 45)
text(x=37, y=0.025, 'Type II error')
# Adding a legend
legend("topright", legend = "X~N(38, 3)", col = "#a1332c",
       title = 'Distribution', cex = 0.75,
       title.adj = 0.9, lty = 1, lwd = 2, box.lty = 0)


# p-value
# Mean 37, sd 3
plot(x, dnorm(x, mean = 37, sd = 3), type = "l", xlab = 'Demand for tickets(in thousands)',
     ylab = '', ylim = c(0, 0.15), lwd = 2, col = "#3a5f6f")
# p-value
segments(43, 0, 43, dnorm(43, 37, 3), lty = 2, lwd = 2, col = '#3a5f6f')
shade_x = c(seq(43,52,0.1), seq(52,43,-0.1))
shade_y = c(dnorm(seq(43,52,0.1), 37, 3), rep(0, length(shade_x)/2))
polygon(x = shade_x, y = shade_y, col = "#3a5f6f", density = 15, angle = 45)
text(x=44, y=0.025, 'p-value')
# Adding a legend
legend("topright", legend = "X~N(37, 3)", col = "#3a5f6f",
       title = 'Distribution', cex = 0.75,
       title.adj = 0.9, lty = 1, lwd = 2, box.lty = 0)

mtext("Normal Distribution of demand for tickets for concerts at the O2",
      side =3, line = -2, outer = TRUE)
```