SDA Group Submission Assignment Assign2

Group Gr18

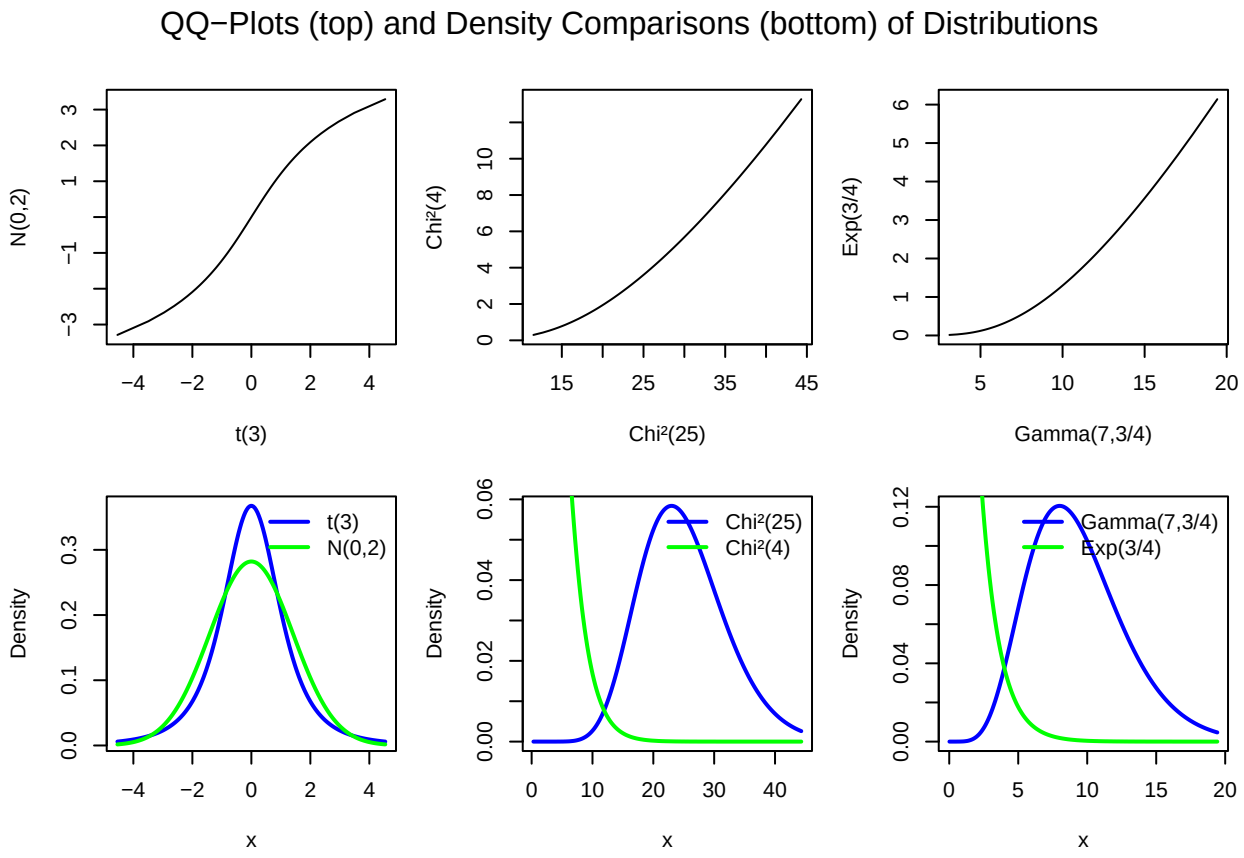
MengliFeng (2720589) and PepijnVanOostveen (2801582)

Exercise 1

a.

```r
# Set up a 2×3 grid for plotting (fills row-wise)
par(mfrow = c(2, 3), mar = c(4, 4, 2, 1), oma = c(0, 0, 2, 0))

# Define probability points
p <- seq(0.01, 0.99, length.out = 100)

# ---- First Row: QQ-Plots ----
# 1. QQ-Plot: t(3) vs Normal(0,2)
q_t3 <- qt(p, df = 3)
q_norm <- qnorm(p, mean = 0, sd = sqrt(2))
plot(q_t3, q_norm, type = "l", xlab = "t(3)", ylab = "N(0,2)")

# 2. QQ-Plot: Chi-squared(25) vs Chi-squared(4)
q_chisq25 <- qchisq(p, df = 25)
q_chisq4 <- qchisq(p, df = 4)
plot(q_chisq25, q_chisq4, type = "l", xlab = "Chi²(25)", ylab = "Chi²(4)")

# 3. QQ-Plot: Gamma(7, 3/4) vs Exponential(3/4)
q_gamma <- qgamma(p, shape = 7, rate = 3/4)
q_exp <- qexp(p, rate = 3/4)
plot(q_gamma, q_exp, type = "l", xlab = "Gamma(7,3/4)", ylab = "Exp(3/4)")

# ---- Second Row: Density Comparison ----
# 4. Density Comparison: t(3) vs N(0,2)
x_vals <- seq(min(q_t3, q_norm), max(q_t3, q_norm), length.out = 200)
plot(x_vals, dt(x_vals, df = 3), type = "l", col = "blue", lwd = 2, xlab = "x", ylab =
↪  "Density", main = "")
lines(x_vals, dnorm(x_vals, mean = 0, sd = sqrt(2)), col = "green", lwd = 2)
legend("topright", legend = c("t(3)", "N(0,2)"), col = c("blue", "green"), lty = 1, lwd =
↪  2, bty = "n")

# 5. Density Comparison: Chi²(25) vs Chi²(4)
x_vals <- seq(min(q_chisq25, q_chisq4), max(q_chisq25, q_chisq4), length.out = 200)
plot(x_vals, dchisq(x_vals, df = 25), type = "l", col = "blue", lwd = 2, xlab = "x", ylab
↪  = "Density", main = "")
lines(x_vals, dchisq(x_vals, df = 4), col = "green", lwd = 2)
legend("topright", legend = c("Chi²(25)", "Chi²(4)"), col = c("blue", "green"), lty = 1,
↪  lwd = 2, bty = "n")
```

```r
# 6. Density Comparison: Gamma(7,3/4) vs Exp(3/4)
x_vals <- seq(min(q_gamma, q_exp), max(q_gamma, q_exp), length.out = 200)
plot(x_vals, dgamma(x_vals, shape = 7, rate = 3/4), type = "l", col = "blue", lwd = 2,
↪  xlab = "x", ylab = "Density", main = "")
lines(x_vals, dexp(x_vals, rate = 3/4), col = "green", lwd = 2)
legend("topright", legend = c("Gamma(7,3/4)", "Exp(3/4)"), col = c("blue", "green"), lty
↪  = 1, lwd = 2, bty = "n")

# --- Add an Overall Title ---
mtext("QQ-Plots (top) and Density Comparisons (bottom) of Distributions", outer = TRUE,
↪  cex = 1, font = 0.5)
```



b.

```r
# Load necessary package
library(EnvStats)
```

```
##
## Attaching package: 'EnvStats'
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```r
# Read the data
mystery <- readRDS("mysterious_sample.RDS")

# Compute sample statistics
y_bar <- mean(mystery)  # Sample mean
s_y <- sd(mystery)  # Sample standard deviation

# Define theoretical distribution parameters
# 1. Uniform(0,1)
mu_x_unif <- (0 + 1) / 2  # Mean of Uniform(0,1)
s_x_unif <- sqrt((1 - 0)^2 / 12)  # Standard deviation of Uniform(0,1)
b_hat_unif <- s_y / s_x_unif
a_hat_unif <- y_bar - b_hat_unif * mu_x_unif

# 2. Extreme Value Distribution (EVD(0,1))
mu_x_evd <- 0.5772  # Mean of EVD(0,1)
s_x_evd <- sqrt(pi^2 / 6)  # Standard deviation of EVD(0,1)
b_hat_evd <- s_y / s_x_evd
a_hat_evd <- y_bar - b_hat_evd * mu_x_evd

# 3. Logistic(0,1)
mu_x_logis <- 0  # Mean of Logistic(0,1)
s_x_logis <- sqrt(pi^2 / 3)  # Standard deviation of Logistic(0,1)
b_hat_logis <- s_y / s_x_logis
a_hat_logis <- y_bar - b_hat_logis * mu_x_logis

# Print estimated slopes and intercepts
cat("Estimated for Uniform(0,1): a =", a_hat_unif, ", b =", b_hat_unif, "\n")
```

```
## Estimated for Uniform(0,1): a = -1.734457 , b = 14.00126
```

```r
cat("Estimated for Extreme Value(0,1): a =", a_hat_evd, ", b =", b_hat_evd, "\n")
```

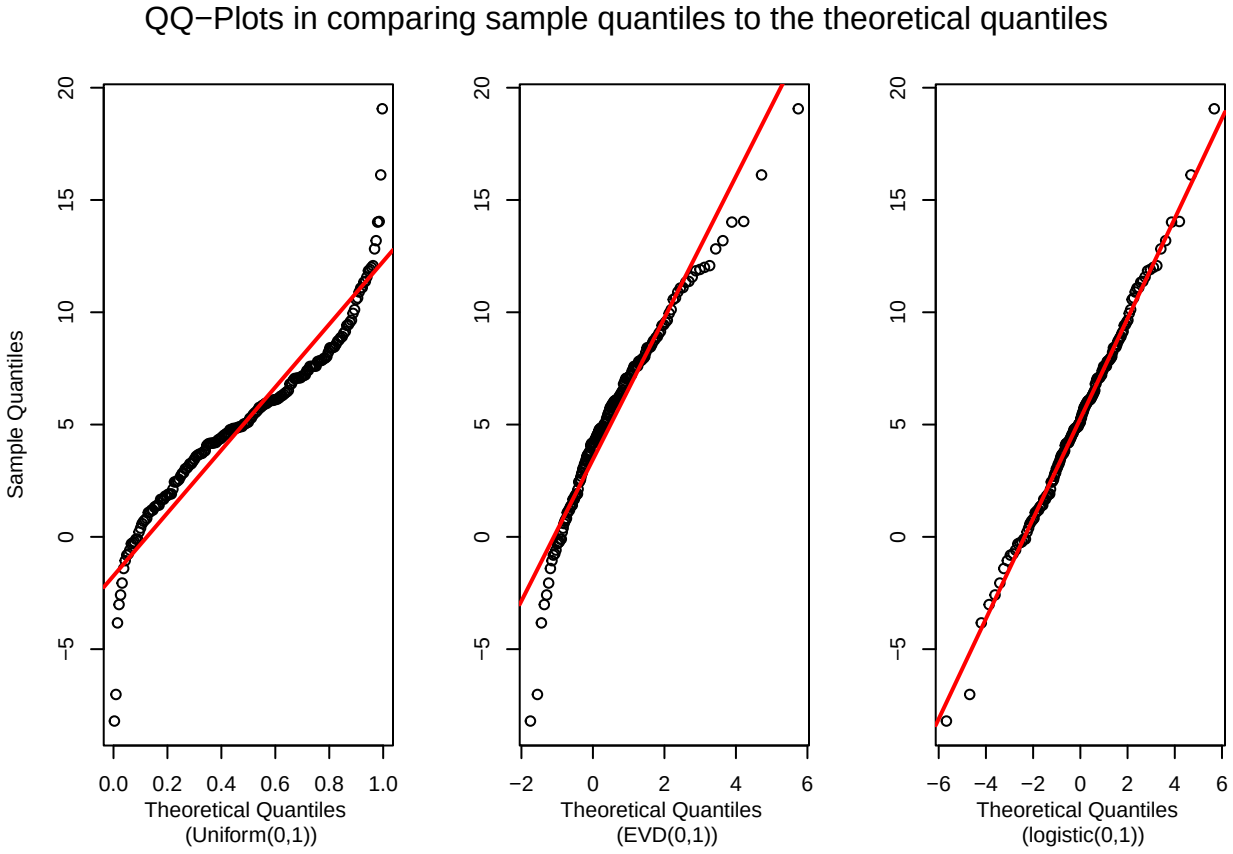
```
## Estimated for Extreme Value(0,1): a = 3.447191 , b = 3.151391
```

```r
cat("Estimated for Logistic(0,1): a =", a_hat_logis, ", b =", b_hat_logis, "\n")
```

```
## Estimated for Logistic(0,1): a = 5.266175 , b = 2.22837
```

```r
# Set up a 1x3 grid for plotting
par(mfrow = c(1, 3), mar = c(4, 4, 2, 1), oma = c(0, 0, 2, 0))

# Generate QQ-Plots with Fitted Lines

# 1. QQ-Plot: Sample vs Uniform(0,1)
qqPlot(mystery, distribution = "unif",
       param.list = list(min = 0, max = 1),
       add.line = FALSE, main = "",
       xlab = "Theoretical Quantiles \n (Uniform(0,1))",
       ylab = "Sample Quantiles")
abline(a = a_hat_unif, b = b_hat_unif, col = "red", lwd = 2)  # Add fitted line

# 2. QQ-Plot: Sample vs Extreme Value(0,1)
qqPlot(mystery, distribution = "evd",
       param.list = list(location = 0, scale = 1),
```

```
       add.line = FALSE, main = "",
       xlab = "Theoretical Quantiles \n (EVD(0,1))",
       ylab = " ")
abline(a = a_hat_evd, b = b_hat_evd, col = "red", lwd = 2)  # Add fitted line

# 3. QQ-Plot: Sample vs Logistic(0,1)
qqPlot(mystery, distribution = "logis",
       param.list = list(location = 0, scale = 1),
       add.line = FALSE, main = "",
       ylab = " ",
       xlab = "Theoretical Quantiles \n (logistic(0,1))")
abline(a = a_hat_logis, b = b_hat_logis, col = "red", lwd = 2)  # Add fitted line

# --- Add an Overall Title ---
mtext("QQ-Plots in comparing sample quantiles to the theoretical quantiles", outer =
↪  TRUE, cex = 1, font = 0.5)
```

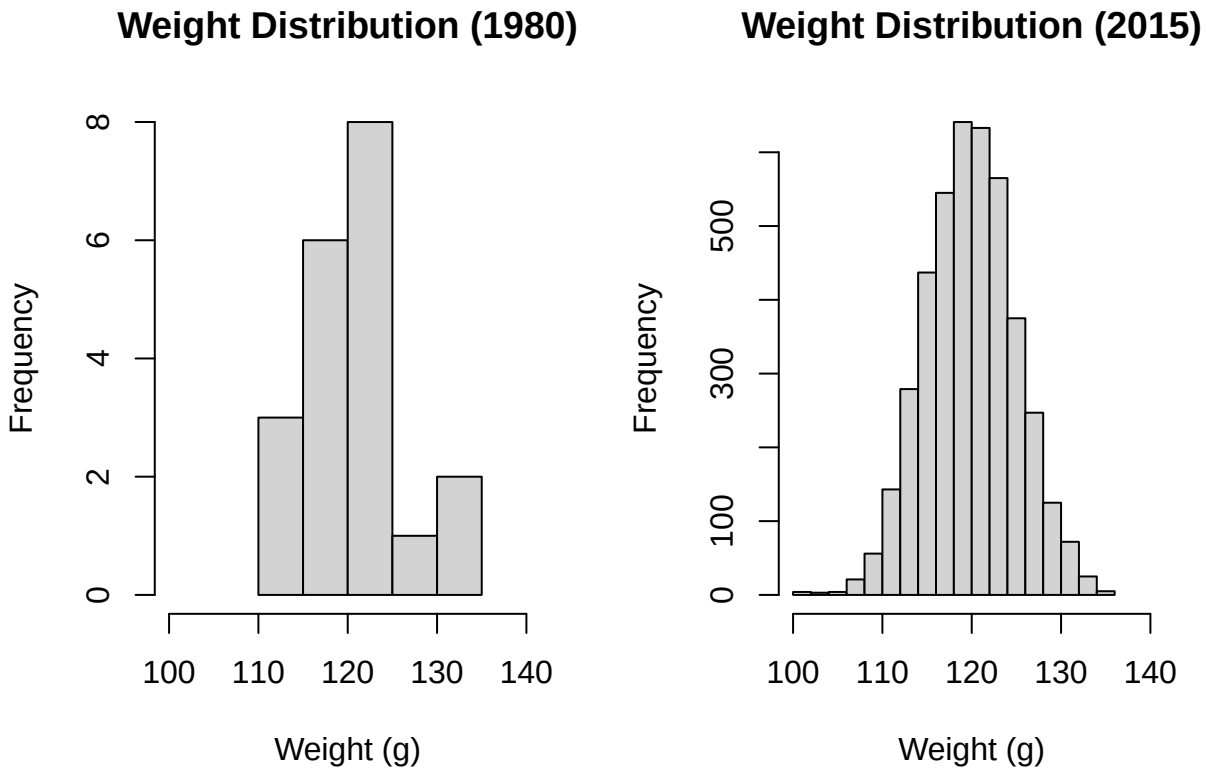


In the plot, the black circles are points of sample quantiles plotted against theoretical quantiles, and the red lines are fitted line given those points with parameter a (intercept) and b (slope). As can be seen from the qq-plots, the logistic distribution fits the data the best because the qq-plot almost looks like a straight line. the other two qq-plots are more curvy and have tails. Estimated for Uniform(0,1): a = -1.734457 , b = 14.00126 Estimated for Extreme Value(0,1): a = 3.447191 , b = 3.151391 Estimated for Logistic(0,1): a = 5.266175 , b = 2.22837

Exercise 2

a.

```r
# load dataset
parakeets <- readRDS("parakeets.RDS")

# select the list of weights where the year is 1980 and 2015 respectively
parakeet_weights_1980 <- parakeets$Weight[parakeets$Year == 1980]
parakeet_weights_2015 <- parakeets$Weight[parakeets$Year == 2015]

# set the grid to 1x2 for plotting the histograms
par(mfrow = c(1, 2))

# plotting for 1980
hist(parakeet_weights_1980, xlim = c(100, 140), main = "Weight Distribution (1980)",
     xlab = "Weight (g)")

# plotting for 2015
hist(parakeet_weights_2015, xlim = c(100, 140), main = "Weight Distribution (2015)",
     xlab = "Weight (g)")
```

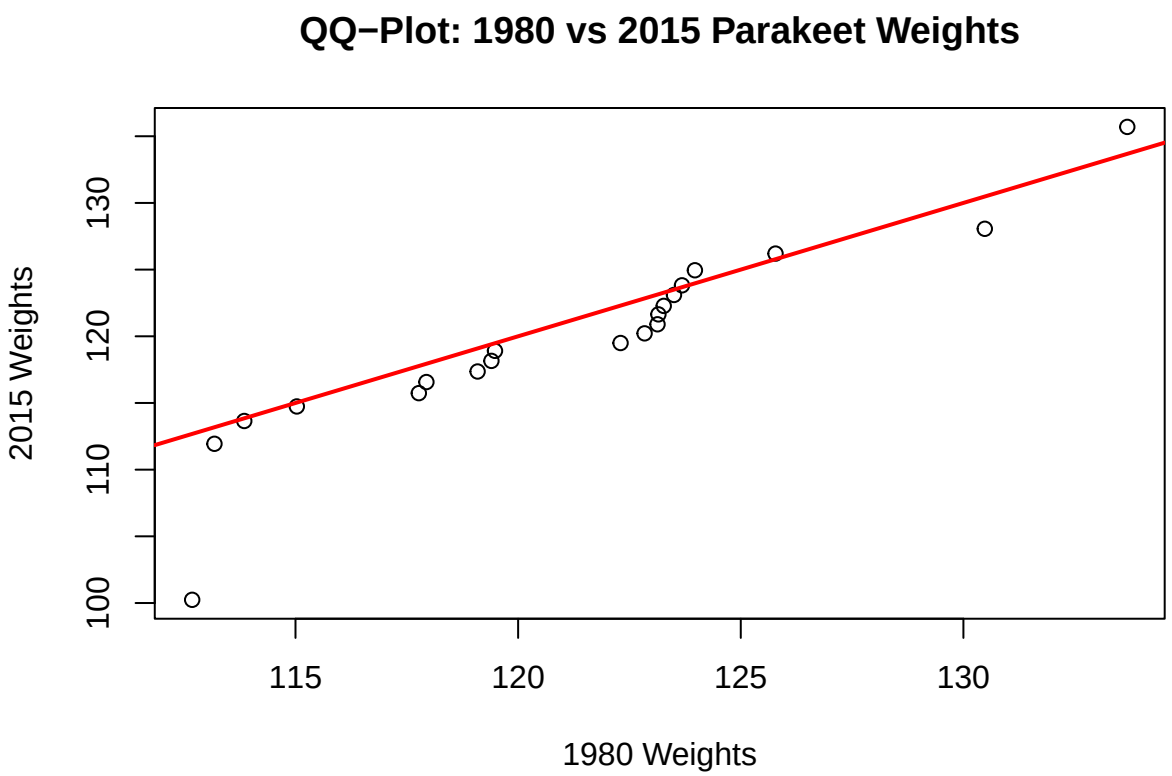


The underlying distributions could be the same since the peak is at around the same weight, but this is hard to say since 1980 only has 20 samples.

b.

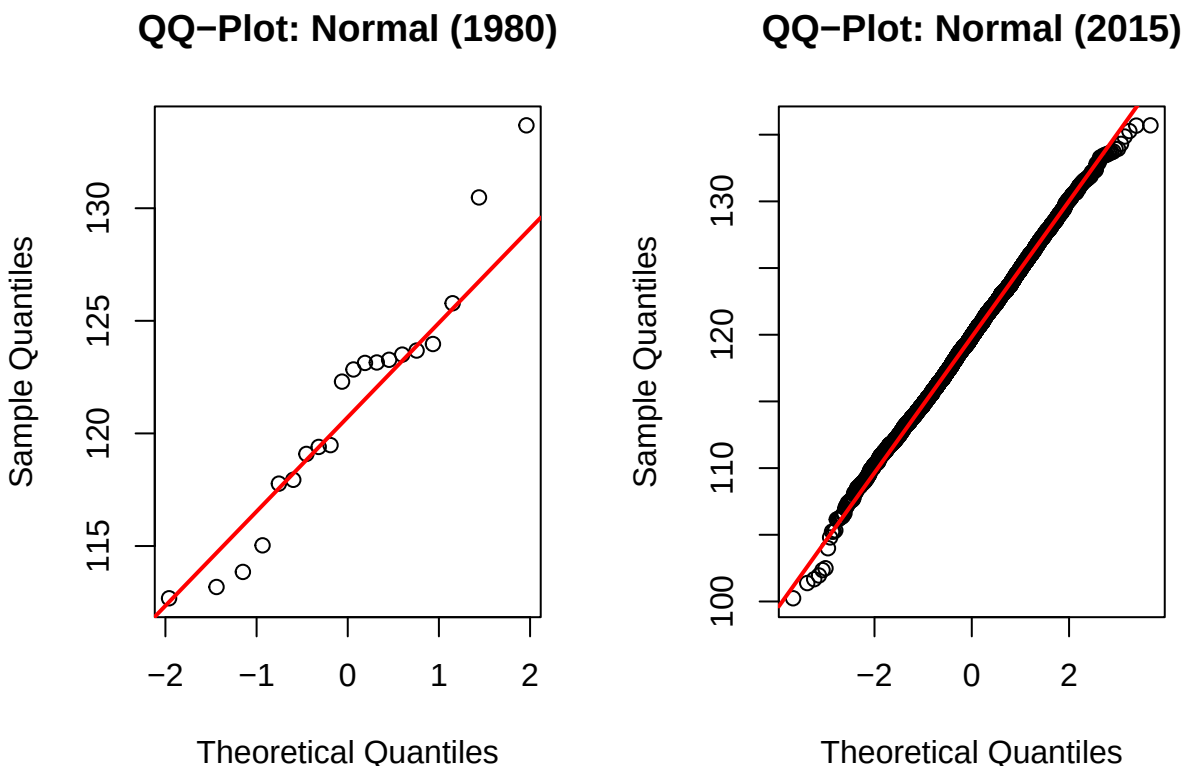
```r
qqplot(parakeet_weights_1980, parakeet_weights_2015,
       main = "QQ-Plot: 1980 vs 2015 Parakeet Weights",
```

```
       xlab = "1980 Weights",
       ylab = "2015 Weights")

# Add a reference line
abline(0, 1, col = "red", lwd = 2)
```



The points approximately follow the reference line so the distributions have the same shape. Thus they could be the same location-scale family (looking at the 2015 weight histogram they are probably both normal distributions). ## c.

```
# set the grid to 1x2 for plotting the QQ-plots
par(mfrow = c(1,2))

# QQ-Plot for 1980 Weights
qqnorm(parakeet_weights_1980, main = "QQ-Plot: Normal (1980)")
# Add a reference line
qqline(parakeet_weights_1980, col = "red", lwd = 2)

# QQ-Plot for 2015 Weights
qqnorm(parakeet_weights_2015, main = "QQ-Plot: Normal (2015)")
# Add a reference line
qqline(parakeet_weights_2015, col = "red", lwd = 2)
```



The weights from the 2015 sample seem to very clearly follow a normal distribution. The weights from the 1980 sample seem the deviate a little more from the fitted line, but they still follow it so it probably still originates from a normal distribution. The 1980 sample needs more weights to get more confidence in this conclusion.

d.

```
shapiro_1980 <- shapiro.test(parakeet_weights_1980)
shapiro_2015 <- shapiro.test(parakeet_weights_2015)

print(shapiro_1980)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  parakeet_weights_1980
## W = 0.94612, p-value = 0.312
```

```
print(shapiro_2015)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  parakeet_weights_2015
## W = 0.99919, p-value = 0.0496
```

The Null hypothesis isn't rejected for the 1980 sample since the p-value is more than 0.05 (corresponding to the significance level of 5%) which doesn't mean that it is normally distributed but it says that we don't have significant proof that it isn't normally distributed. The Null hypothesis for the the 2015 sample is rejected since the p-value of 0.0496 is lower than 0.05. This means we have significant proof that the 2015 sample isn't normally distributed.

e.

In the graphs for the 1980 sample we couldn't say for sure that it followed a normal distribution, but it looked like it did and the Shapiro-Wilk normality test said that it is either normal or we don't have enough data to be confident that it isn't normal. Thus the conclusion is that the 1980 sample is probably follows a normal distribution but it is inconclusive because of the small sample size. The graphs for the 2015 sample show that it follows a normal distribution. The Shapiro-Wilk normality test says that the sample isn't normal, but this test is more sensitive to small deviations from a normal distribution as the sample size increases and from the graphs you can see that this probably happens due to the tails that deviate from a normal distribution. Thus the 2015 sample doesn't follow a normal distribution exactly, but the distribution that it follows can be approximated really well by a normal distribution, especially in the area exclusing the tails.