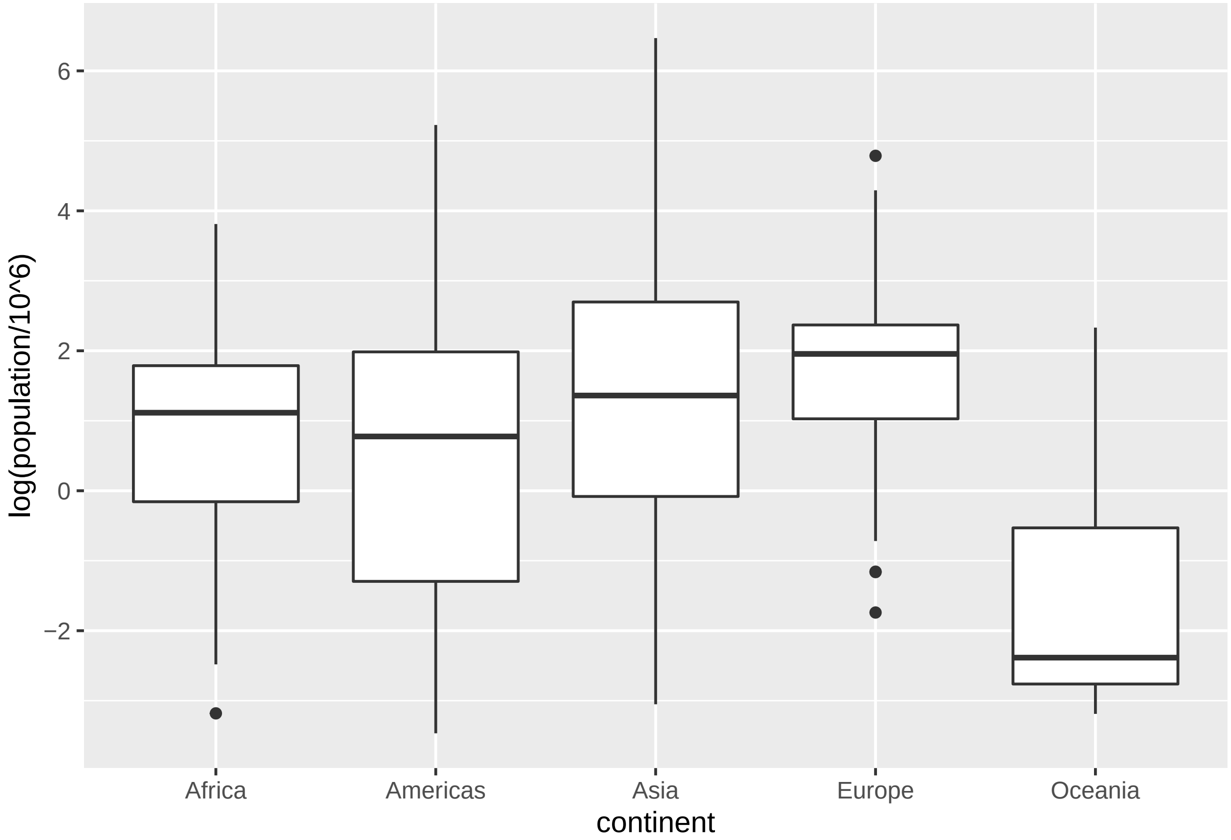
```
rm(list=ls())

library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
load(file = "country.rda")
```

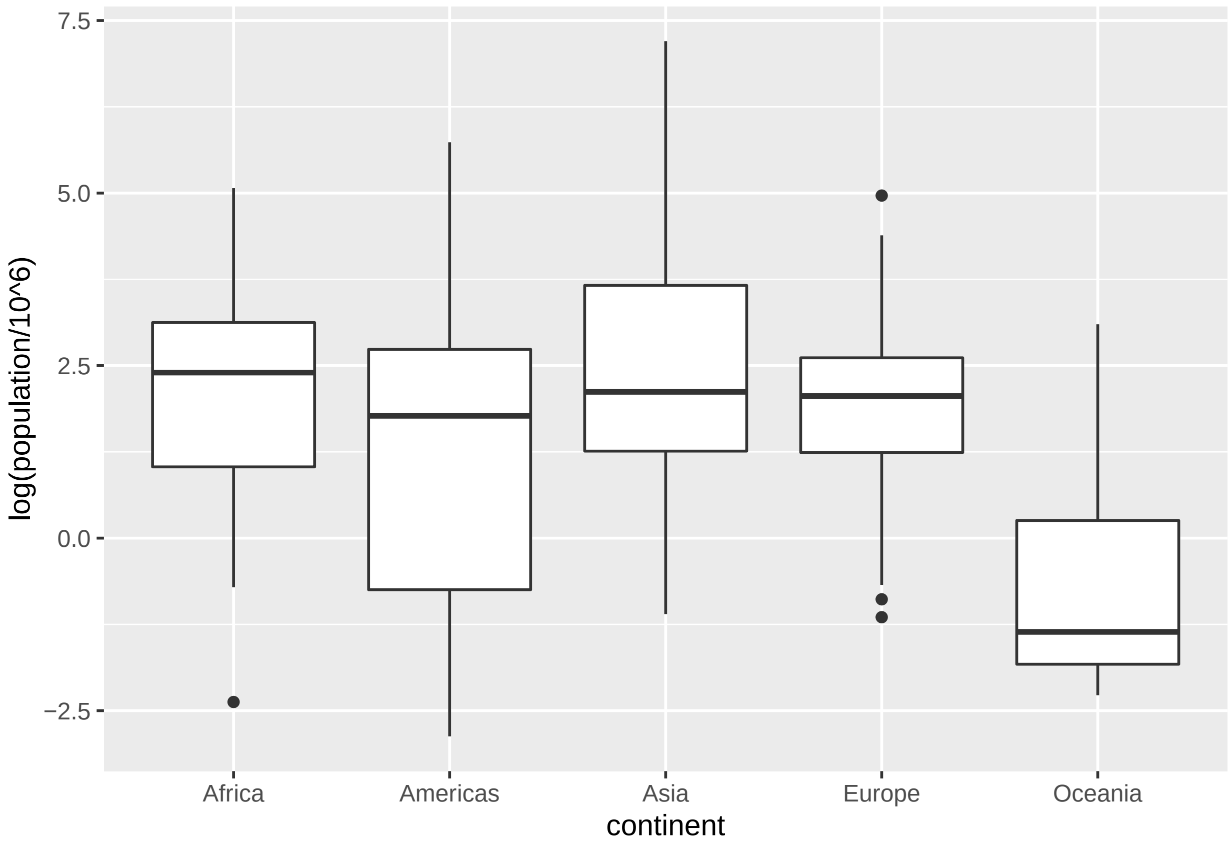
(1) Solution: Europe. Africa, Americas, Asia and Europe all have close median population size from the plot. Oceania's median population size is extrememly low because there are few countries in Oceania, causing some bias from the model of previous 4 continents.

```
df_1 <- gapminder[gapminder$year == 1960,]

p1 <- ggplot(data = df_1, aes(x=continent, y=log(population/10^6))) + geom_boxplot()
p1
```



(2) Solution: Africa. 1960 and 2010 have some difference according to the specific figure about each continent's population size but the trend is almost the same across two years.

```
df_2 <- gapminder[gapminder$year == 2010,]
p2 <- ggplot(data = df_2, aes(x=continent, y=log(population/10^6))) + geom_boxplot()
p2
```



(3) Solution: The nearest million is 6 million. 0.7566.

```
df_3_Africa <- gapminder[gapminder$continent == 'Africa',]
round(median(df_3_Africa$population, na.rm = TRUE), -6)
```

```
## [1] 6e+06
```

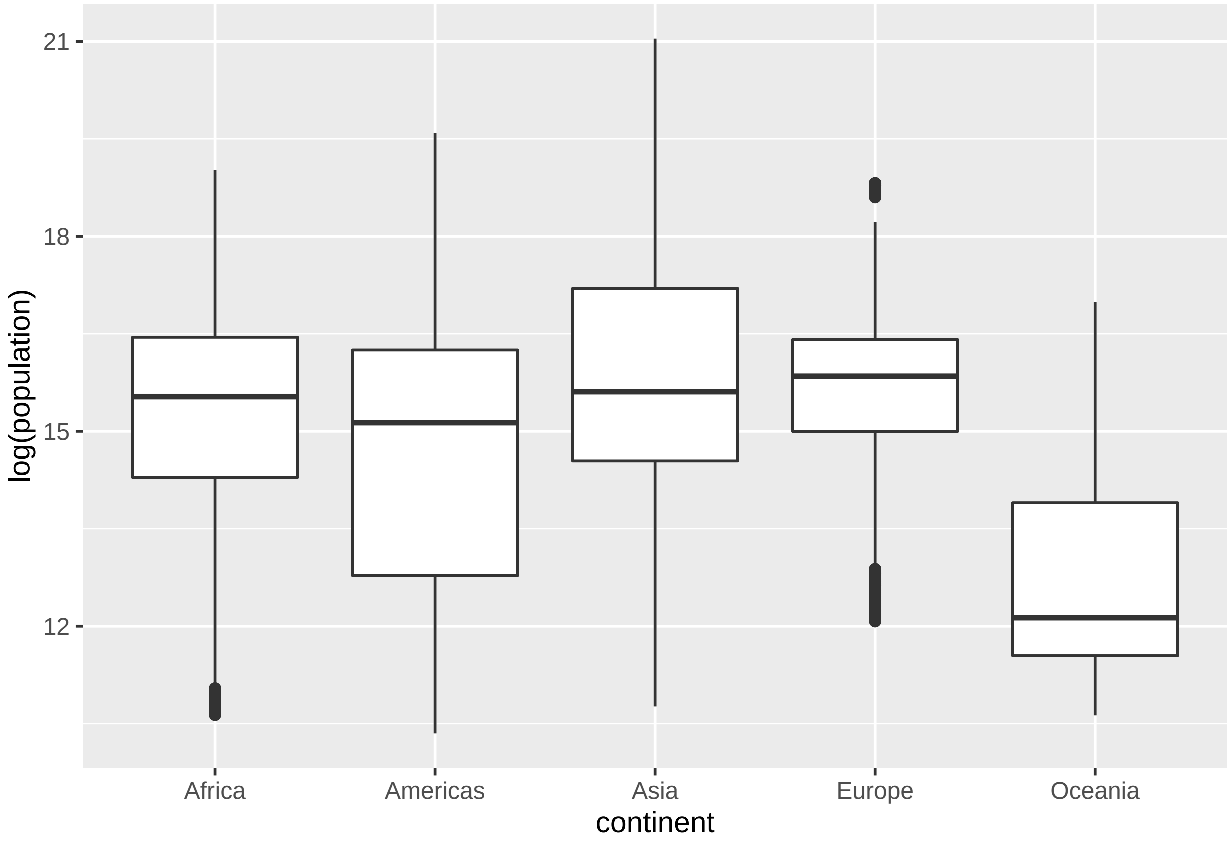
```
df_3_Euro = gapminder[gapminder$continent == 'Europe',]
df_3_res <- df_3_Euro[df_3_Euro$population < 14000000, ]$country
length(df_3_res)/nrow(df_3_Euro)
```

```
## [1] 0.7566352
```

(4) Solution: The boxplot indicates that the 'Americas' has the largest interquartile range. The size of the boxes in the boxplot represents the interquartile for each continent. 'Americas' has the biggest box.

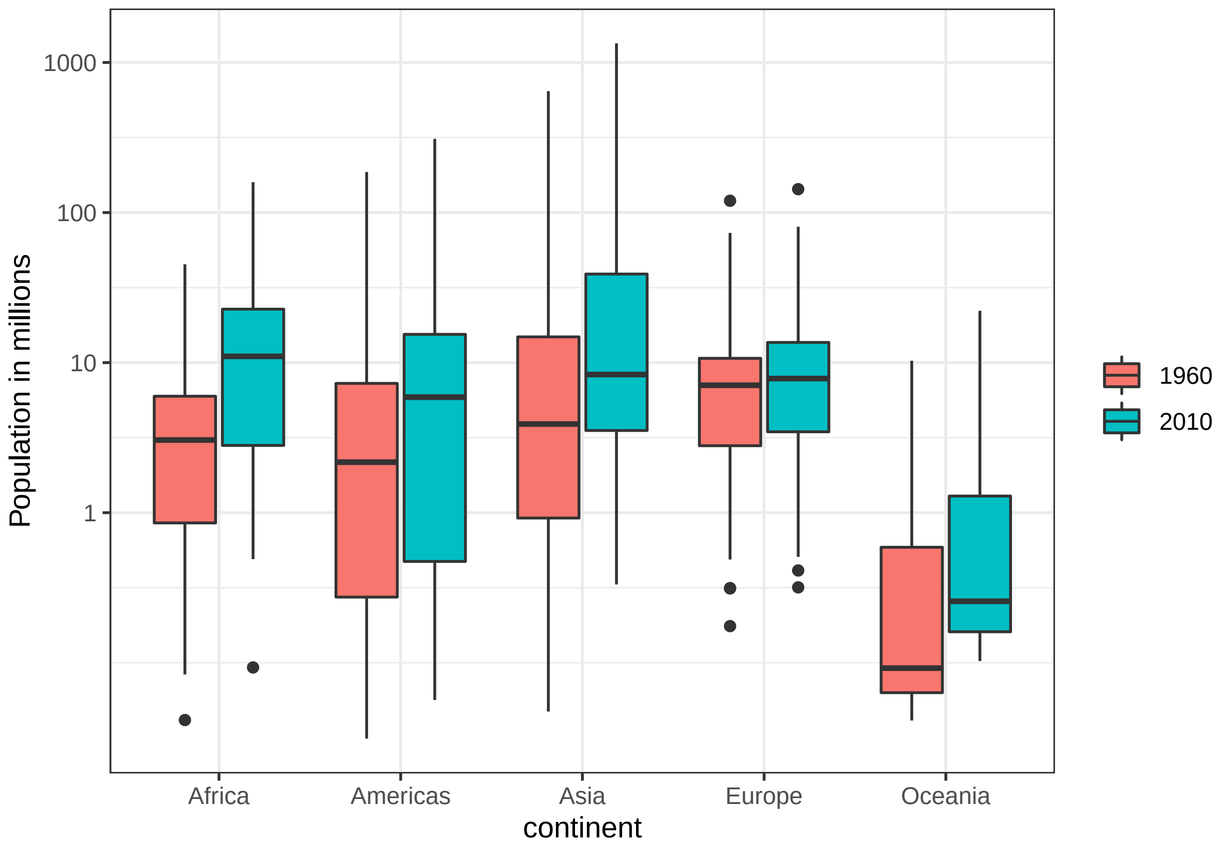
```
p4 <- ggplot(data = gapminder, aes(x=continent, y=log(population))) + geom_boxplot()
p4
```

```
## Warning: Removed 185 rows containing non-finite values (stat_boxplot).
```



(5)

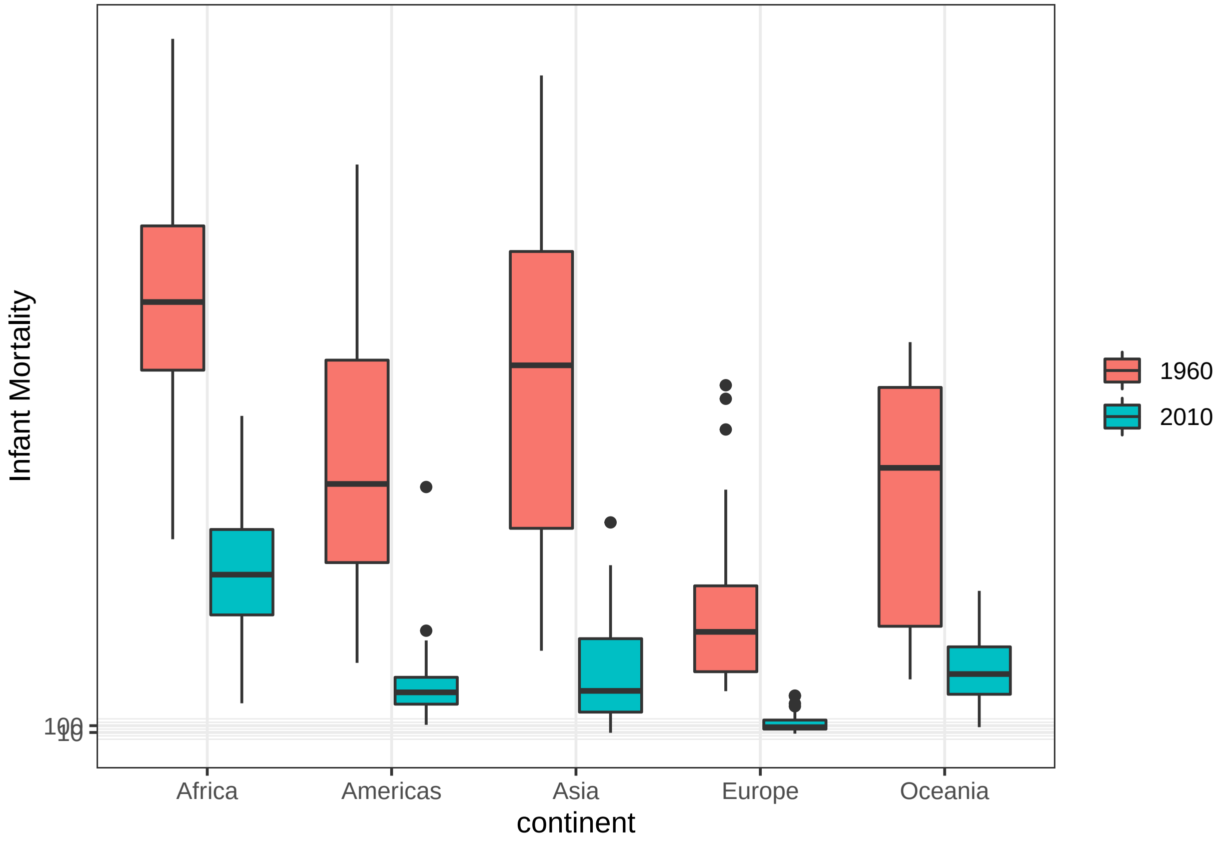
```
df_5_pop <- gapminder[gapminder$year %in% c(1960,2010),]
p5 <-ggplot(df_5_pop, aes(x=continent, y=log(population/10^6), fill=as.factor(year))) +
  geom_boxplot()+
  theme_bw() +
  scale_y_continuous(name="Population in millions", breaks = c(log(1),log(10),log(100),log(1000)), label
  theme(legend.title = element_blank())
p5
```



(6)

```
df_6_infant <- gapminder[gapminder$year %in% c(1960,2010),]
p6 <-ggplot(df_6_infant, aes(x=continent, y=infant_mortality, fill=as.factor(year))) +
  geom_boxplot() +
  theme_bw() +
  scale_y_continuous(name="Infant Mortality", breaks = c(log(10),log(100)), labels= c(10, 100)) +
  theme(legend.title = element_blank())
p6
```

```
## Warning: Removed 52 rows containing non-finite values (stat_boxplot).
```



(7) Solutions: I assume a normal distribution on this dataset. 0.1257 between 100 and 150, and the approximation proportion is 0.1509 in normal distribution. 0.0391 is greater than 150, and the approximation proportion is 0.0236 in normal distribution.

```
df_7_1 <-  gapminder %>% filter(infant_mortality > 100 & infant_mortality<=150)
nrow(df_7_1) / nrow(gapminder)
```

```
## [1] 0.1257468
```

```
df_7_2 <-  gapminder %>% filter(infant_mortality > 150)
nrow(df_7_2) / nrow(gapminder)
```

```
## [1] 0.03907065
```

```
mean_7 = mean(gapminder$infant_mortality, na.rm = TRUE)
std_7 = sd(gapminder$infant_mortality, na.rm = TRUE)
pnorm(150, mean_7, std_7) - pnorm(100, mean_7, std_7)
```

```
## [1] 0.1509108
```

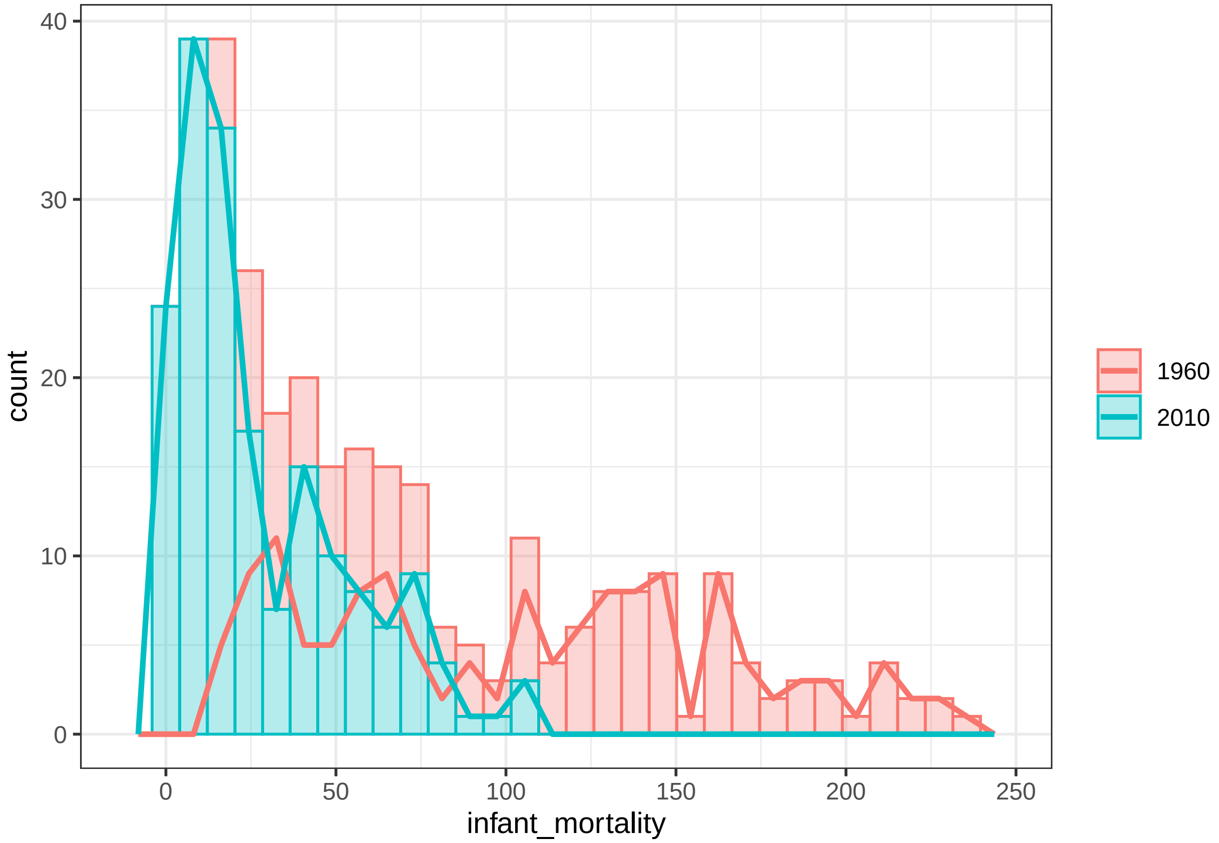
```
1- pnorm(150, mean_7, std_7)
```

```
## [1] 0.02362918
```

(8) Solution: Both of them are not normal distributions.

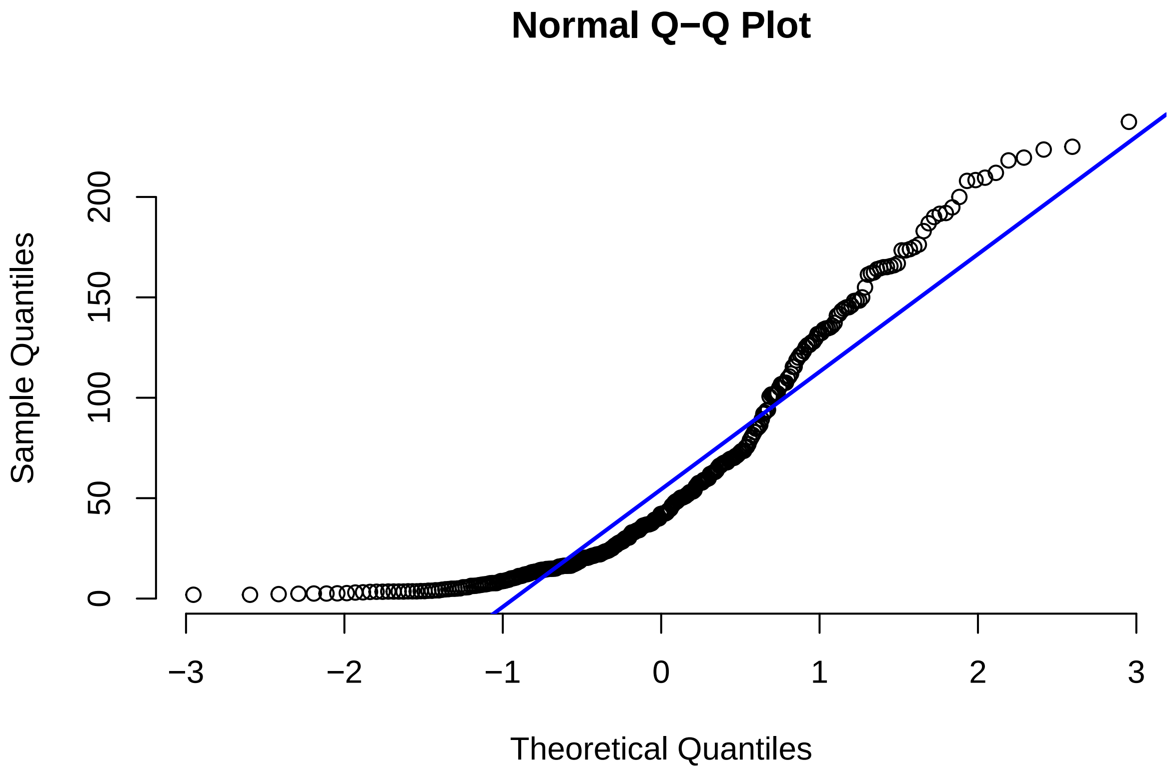
```
p8 <- ggplot(df_6_infant, aes(infant_mortality, fill = as.factor(year), colour = as.factor(year))) +
  geom_histogram(alpha = 0.3) +
  geom_freqpoly(size=1) +
  theme_bw() +
  theme(legend.title = element_blank())
p8
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 52 rows containing non-finite values (stat_bin).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 52 rows containing non-finite values (stat_bin).
```



(9) Solution: If sample_quantiles could fit theoretical_quantiles, it means the sample strictly follows the assmuming model. In this plot, the points are around the fitted line, which means the sample is very closed to the normal distribution. p9 <- seq(0.05, 0.95, 0.05) sample_quantiles <- quantile(df_infant$infant_{m}ortality, p9, na.rm = TRUE)theoretical_{q}uantiles < -qnorm(p9, mean = mean(df_{i}nfant$infant_mortality,na.rm=TRUE), sd = sd(df_infant$infant_mortality, na.rm=TRUE)) qplot(theoretical_quantiles, sample_quantiles) + geom_abline()

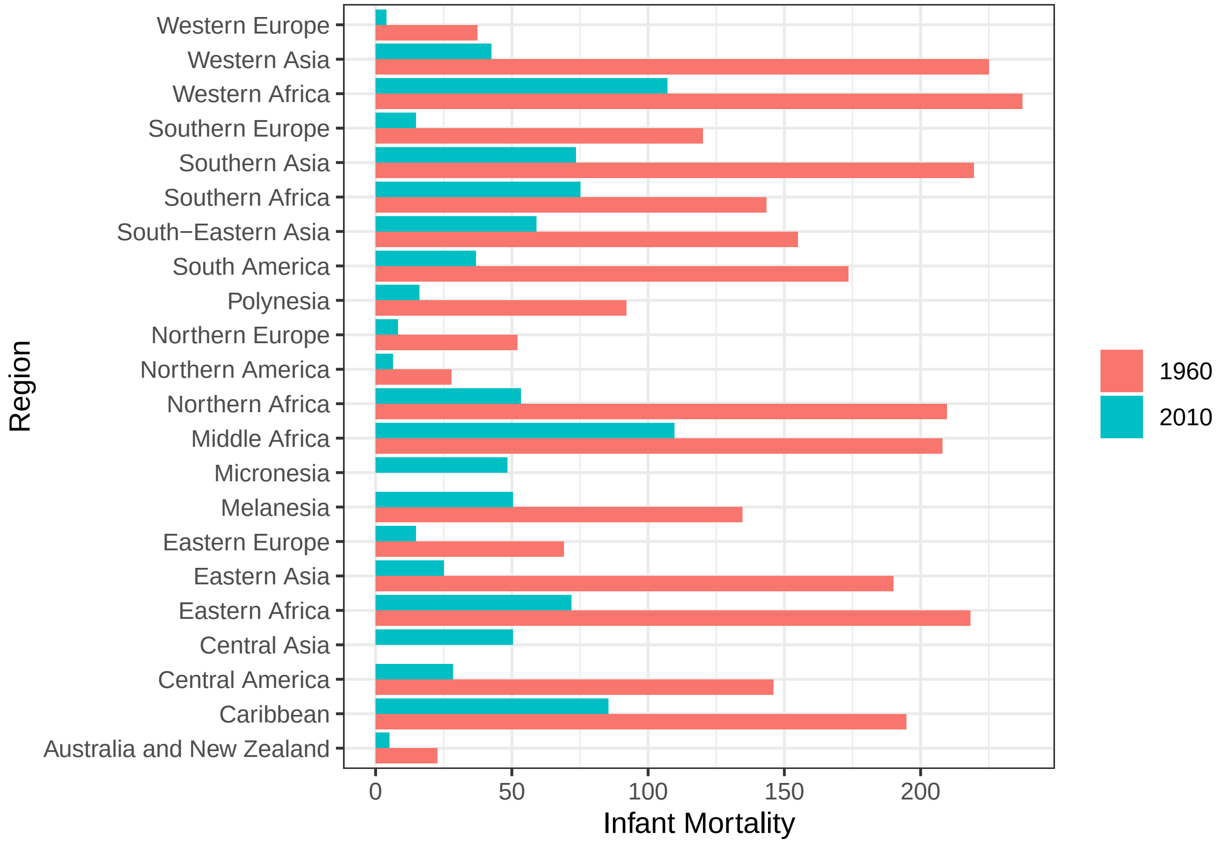
```
x <- df_6_infant$infant_mortality
qqnorm(x, pch =1, frame = FALSE)
qqline(x, col = "blue", lwd = 2)
```



(10)

```
p10 <- ggplot(data=df_6_infant, aes(x=region, y=infant_mortality, fill=as.factor(year))) +
  geom_bar(stat="identity", position=position_dodge()) +
  coord_flip() +
  theme_bw() +
  labs(y = "Infant Mortality", x = "Region") +
  theme(legend.title = element_blank())
p10
```

```
## Warning: Removed 52 rows containing missing values (geom_bar).
```


Western Europe
Western Asia
Western Africa
Southern Europe
Southern Asia
Southern Africa
South−Eastern Asia
South America
Polynesia
Northern Europe
Northern America
Northern Africa
Middle Africa
Micronesia
Melanesia
Eastern Europe
Eastern Asia
Eastern Africa
Central Asia
Central America
Caribbean
Australia and New Zealand
Region
0
50
100
150
200
Infant Mortality
1960
2010