Q0: Sampling and Distribution

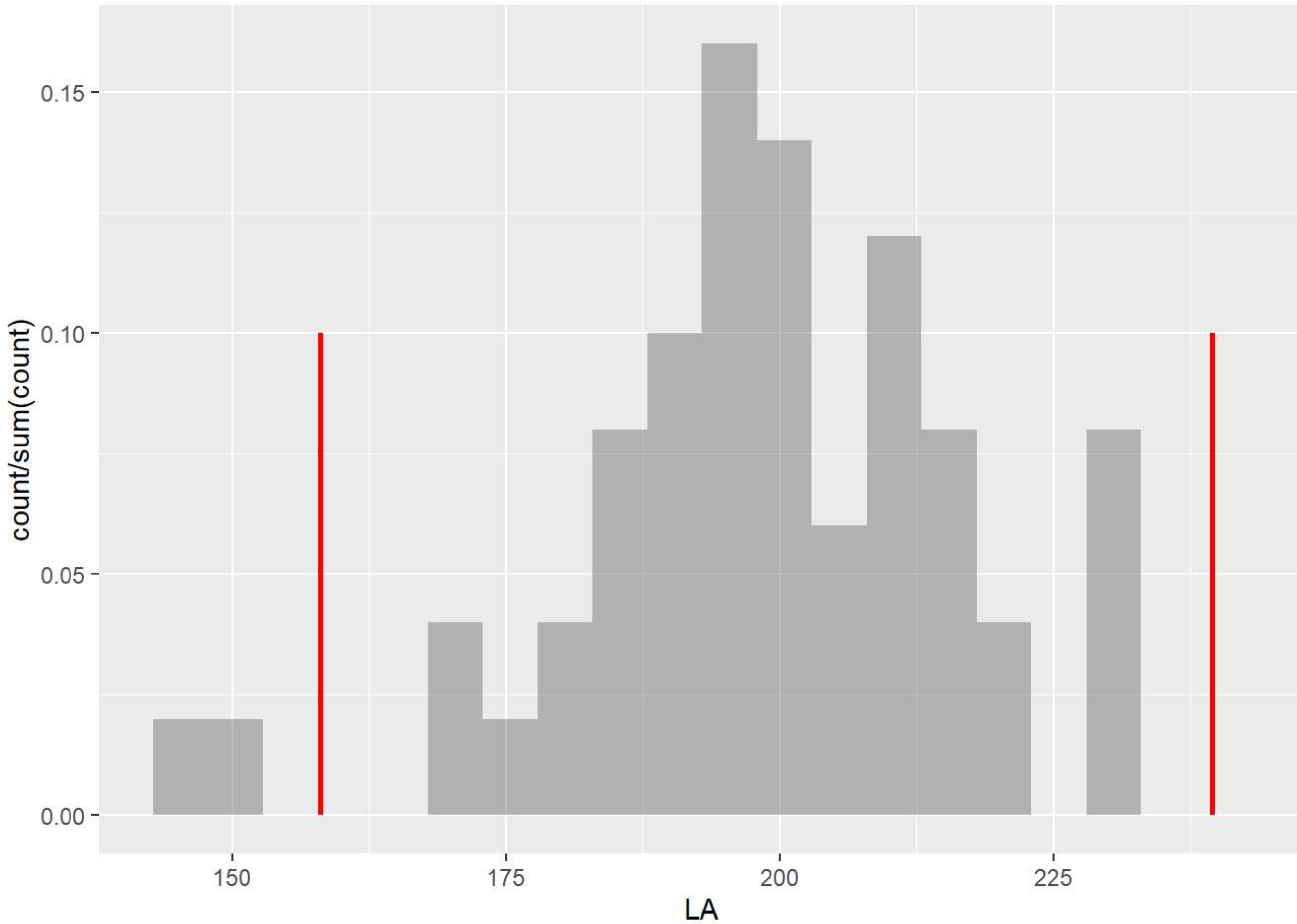
You are studying the population of a specific type of marine algae in different locations. Assume the algal density is normally distributed. You take samples from two locations (Location A and Location B) to compare the algae populations. Generate synthetic data to represent the algal density (individuals per square meter) at these two locations. Assume a mean density of 200 and 220 individuals/m^2 with a common standard deviation of 20 individuals/m^2 for both locations, with 50 samples from each location.

```
#Location A
LA <- rnorm(50, mean = 200, sd = 20)

#Location B
LB <- rnorm(50, mean = 220, sd = 20)
```

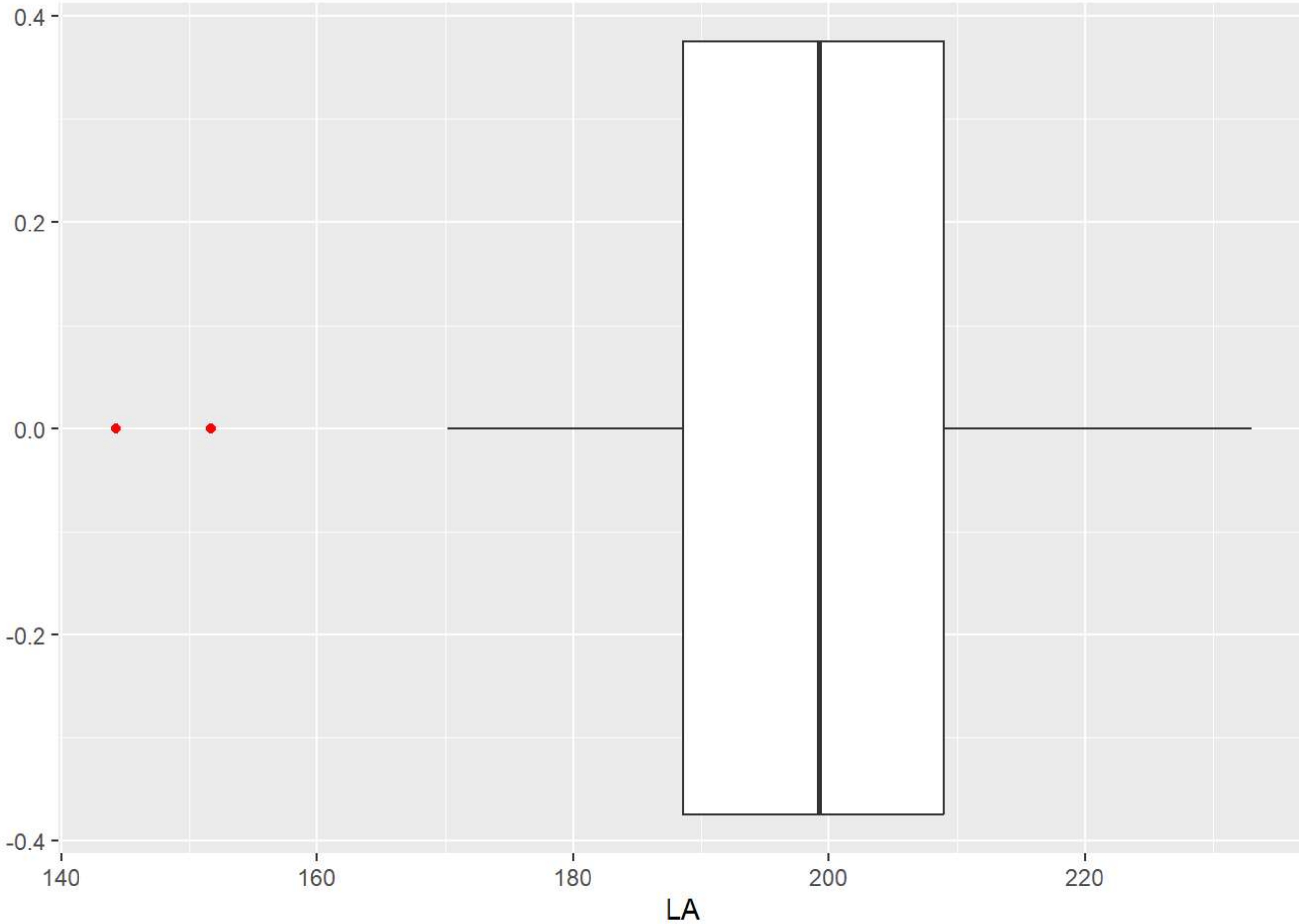
Q1: Data Cleaning

Check your dataset for any outliers.

```
library(ggplot2)
library(patchwork)

Q1 <- summary(LA)[2]
Q3 <- summary(LA)[5]

bound <- (Q3-Q1)*1.5

Low.bound <- summary(LA)[2] - bound
Up.bound <- summary(LA)[5] + bound

ggplot() +
  geom_histogram(aes(x = LA, y = after_stat(count / sum(count))), bins = 20, alpha = 0.4) +
  geom_segment(aes(x = Low.bound, y = 0, xend = Low.bound, yend = .10), color="red", linewidth =
1) +
  geom_segment(aes(x = Up.bound, y = 0, xend = Up.bound, yend = .10), color="red", linewidth =1)
```




```
ggplot() +
  geom_boxplot(aes(x = LA), outlier.color = "red")
```

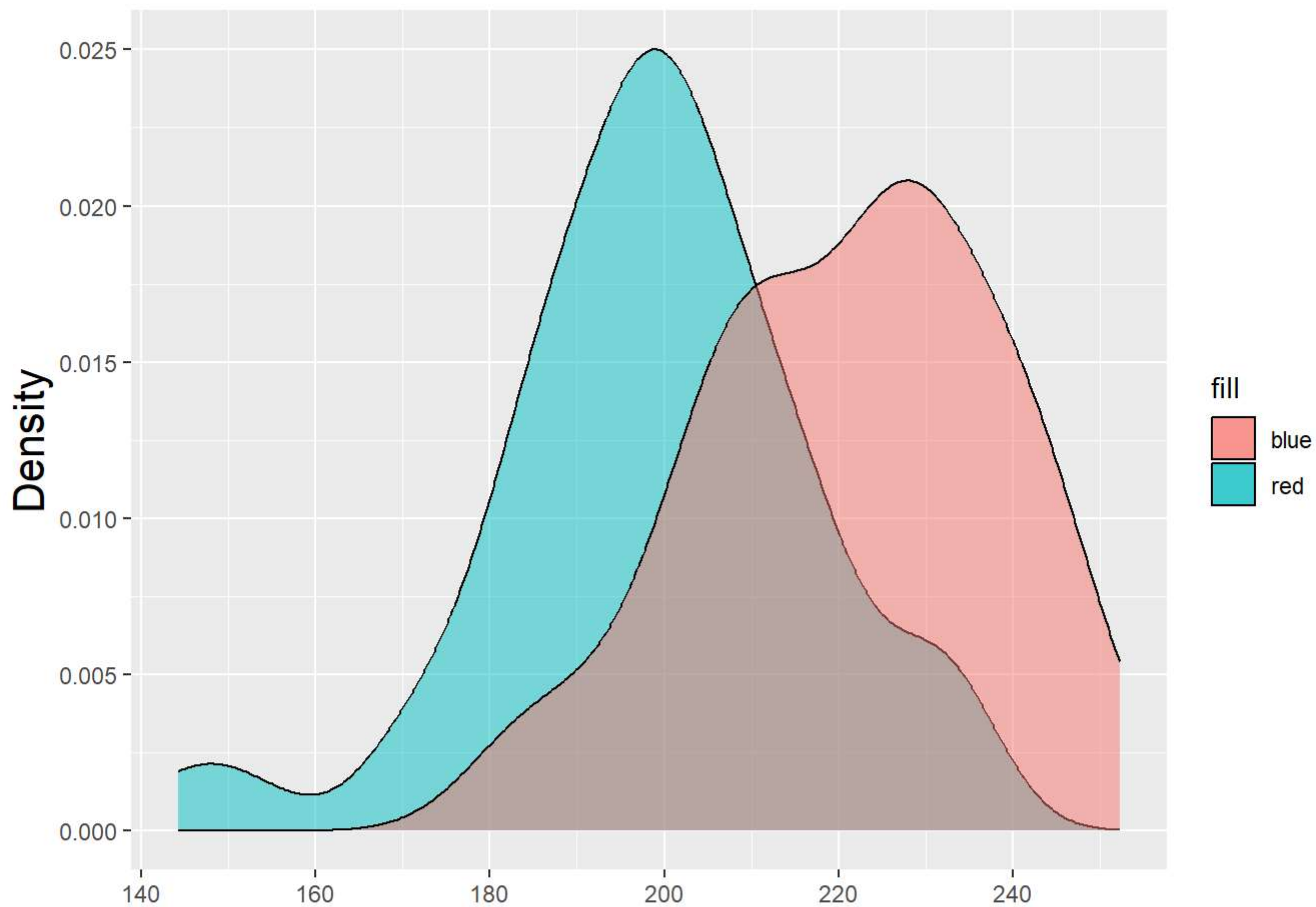


Data positioned above/below the upper/lower bounds of the interquartile range can be considered outliers.

Q2: Visualization and Kernel Density Estimation (KDE)

Plot a Kernel Density Estimation (geom_density plot) to visualize the distribution of algal densities at both locations.

```
ggplot() +
  geom_density(aes(x = LA, fill = "red"), linewidth = 0.5, alpha = 0.5) +
  geom_density(aes(x = LB, fill = "blue"), linewidth = 0.5, alpha = 0.5) +
  scale_color_manual(labels = c("A", "B"), values = c("red", "blue")) +
  theme( axis.title.y = element_text(size = 16)) +
  labs(
    color = "Location",
    x = "",
    y = "Density")
```



Q3: Binomial Distribution

Suppose in a new survey, at each location, you take 10 random samples and in each sample, you identify whether a particular species of marine algae is present or not. Assume the probability of finding this species in a sample is 0.7 at Location A and 0.5 at Location B.

Simulate this scenario using a binomial distribution, and compare the probability of finding the species in at least 7 out of 10 samples at both locations.

```
#Location A
LA_binom <- dbinom(x = 7, size = 10, prob = 0.7)
LA_binom
```

```
## [1] 0.2668279
```

```
#Location B
LB_binom <- dbinom(x = 7, size = 10, prob = 0.5)
LB_binom
```

```
## [1] 0.1171875
```

```
#comparison
((LA_binom - LB_binom)/LB_binom)*100
```

```
## [1] 127.6932
```

The probability of the marine algae species being present in at least 7 out of 10 samples in Location A is 0.2668279; in Location B, it is 0.1171875. Thus, the probability of finding the species in at least 7 out of 10 samples at Location A is 127.7% higher than the probability at Location B.

Q4: Poisson Distribution

Imagine a scenario where you are studying the occurrences of a particular rare marine event, such as the sighting of a rare marine species, over a set period at a specified location. Assume the average rate of occurrence is 3 per month.

Utilize a Poisson distribution to calculate the probability of observing exactly 5 occurrences in a month, and the probability of observing 3 or fewer occurrences in a month.

```
#setting x-axis equal to number of days in a month
x_axis <- 0:30

p_x <- dpois(x_axis, 3)

names(p_x) = x_axis
p_x
```

```
##            0            1            2            3            4            5 
## 4.978707e-02 1.493612e-01 2.240418e-01 2.240418e-01 1.680314e-01 1.008188e-01 
##            6            7            8            9           10           11 
## 5.040941e-02 2.160403e-02 8.101512e-03 2.700504e-03 8.101512e-04 2.209503e-04 
##           12           13           14           15           16           17 
## 5.523758e-05 1.274713e-05 2.731529e-06 5.463057e-07 1.024323e-07 1.807629e-08 
##           18           19           20           21           22           23 
## 3.012715e-09 4.756919e-10 7.135379e-11 1.019340e-11 1.390009e-12 1.813055e-13 
##           24           25           26           27           28           29 
## 2.266319e-14 2.719583e-15 3.137980e-16 3.486644e-17 3.735690e-18 3.864507e-19 
##           30 
## 3.864507e-20
```

```
#probability of observing exactly 5 occurrences in a month
p_x["5"]
```

```
##         5 
## 0.1008188
```

```
#probability of observing 3 or fewer occurrences in a month
sum(p_x[1:4])
```

```
## [1] 0.6472319
```

Q5:

consider the following two lists.

```
list1 <- c(44.40, 47.70, 65.59, 50.71, 51.29, 67.15, 54.61, 37.35, 43.13, 45.54, 62.24, 53.60, 5
4.01, 51.11,
           44.44, 67.87, 54.98, 30.33, 57.01, 45.27, 39.32, 47.82, 39.74, 42.71, 43.75, 33.13, 5
8.38, 51.53,
           38.62, 62.54, 54.26, 47.05, 58.95, 58.78, 58.22, 56.89, 55.54, 49.38, 46.94, 46.20, 4
3.05, 47.92,
           37.35, 71.69, 62.08, 38.77, 45.97, 45.33, 57.80, 49.17, 52.53, 49.71, 49.57, 63.69, 4
7.74, 65.16,
           34.51, 55.85, 51.24, 52.16, 53.80, 44.98, 46.67, 39.81, 39.28, 53.04, 54.48, 50.53, 5
9.22, 70.50,
           45.09, 26.91, 60.06, 42.91, 43.12, 60.26, 47.15, 37.79, 51.81, 48.61, 50.06, 53.85, 4
6.29, 56.44,
           47.80, 53.32, 60.97, 54.35, 46.74, 61.49, 59.94, 55.48, 52.39, 43.72, 63.61, 44.00, 7
1.87, 65.33,
           47.64, 39.74)

list2 <- c(44.34, 48.85, 41.30, 39.79, 30.73, 44.32, 33.23, 19.98, 39.30, 58.78, 36.37, 54.12, 2
0.73, 44.17,
           52.79, 49.52, 46.59, 35.39, 32.25, 29.64, 46.76, 30.79, 37.64, 41.16, 72.66, 35.22, 4
8.53, 46.17,
           30.57, 43.93, 66.67, 51.77, 45.62, 38.66, 14.20, 61.97, 23.09, 56.10, 73.64, 23.34, 5
5.53, 41.07,
           21.42, 22.28, 20.98, 37.04, 23.07, 55.32, 76.50, 25.69, 56.82, 56.54, 49.98, 29.87, 4
3.21, 40.79,
           53.44, 39.41, 59.65, 39.38, 60.79, 29.26, 26.10, 93.62, 38.75, 49.47, 54.55, 37.74, 5
2.75, 50.53,
           41.77, 45.98, 44.49, 76.93, 33.88, 28.56, 45.57, 49.66, 51.55, 38.12, 29.05, 63.95, 3
9.76, 32.02,
           41.46, 42.04, 61.65, 46.27, 56.31, 37.51, 48.22, 40.13, 46.42, 31.57, 25.34, 74.96, 5
4.01, 26.23,
           35.83, 27.22)
```

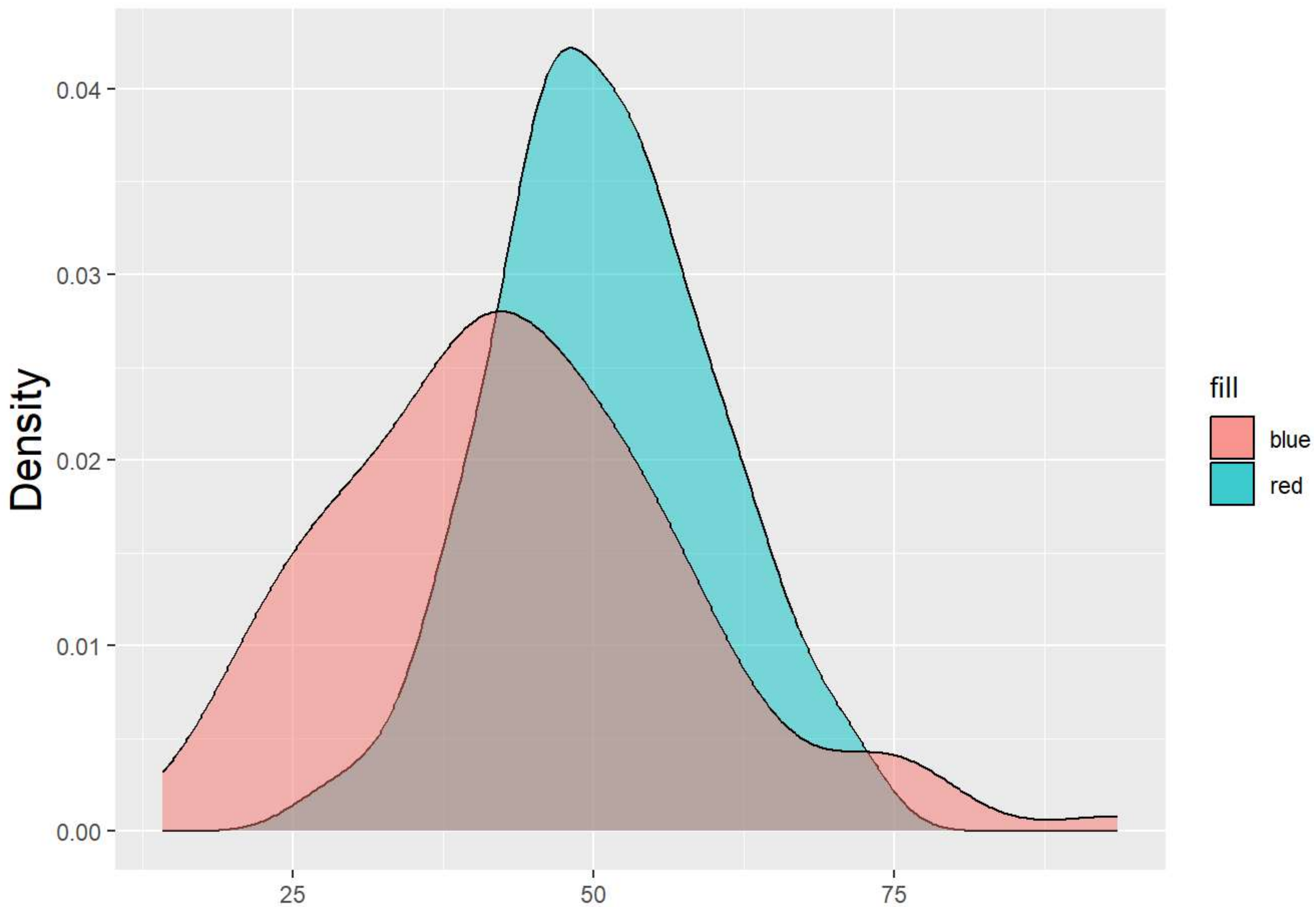
These list were generated using the following code

set.seed(123) # Setting a seed for reproducibility list1 <- round(rnorm(100, mean = 50, sd = 10), 2) # Generating 100 values from a normal distribution list2 <- round(rnorm(100, mean = 45, sd = 15), 2) # Generating 100 values from another normal distribution

Compare the distributions of these two lists to determine if they originate from the same distribution or from different distributions.

1. Use geom_density to create a density plot for each list on the same graph. Use different colors to distinguish between the two lists:

```
ggplot() +
  geom_density(aes(x = list1, fill = "red"), linewidth = 0.5, alpha = 0.5) +
  geom_density(aes(x = list2, fill = "blue"), linewidth = 0.5, alpha = 0.5) +
  scale_color_manual(labels = c("1", "2"), values = c("red", "blue")) +
  theme( axis.title.y = element_text(size = 16)) +
  labs(
    color = "List",
    x = "",
    y = "Density")
```



Examine the plot you have generated. Do you think list1 and list2 come from the same distribution or different distributions? Why? Write down your observations and reasoning.

- Write your answer below

Yes, list1 and list2 both come from a normal/Gaussian distribution, BUT they have different parameters; list1 has a mean of 50 and a sd of 10, but list2 has a mean of 45 and a sd of 15. Additionally, they both appear fairly symmetric and bell-shaped, substantiating that they both likely came from the same type of distribution (normal).