Math 185 Homework 3

Lucien Chen

2024-04-26

Homework 3

Question 1

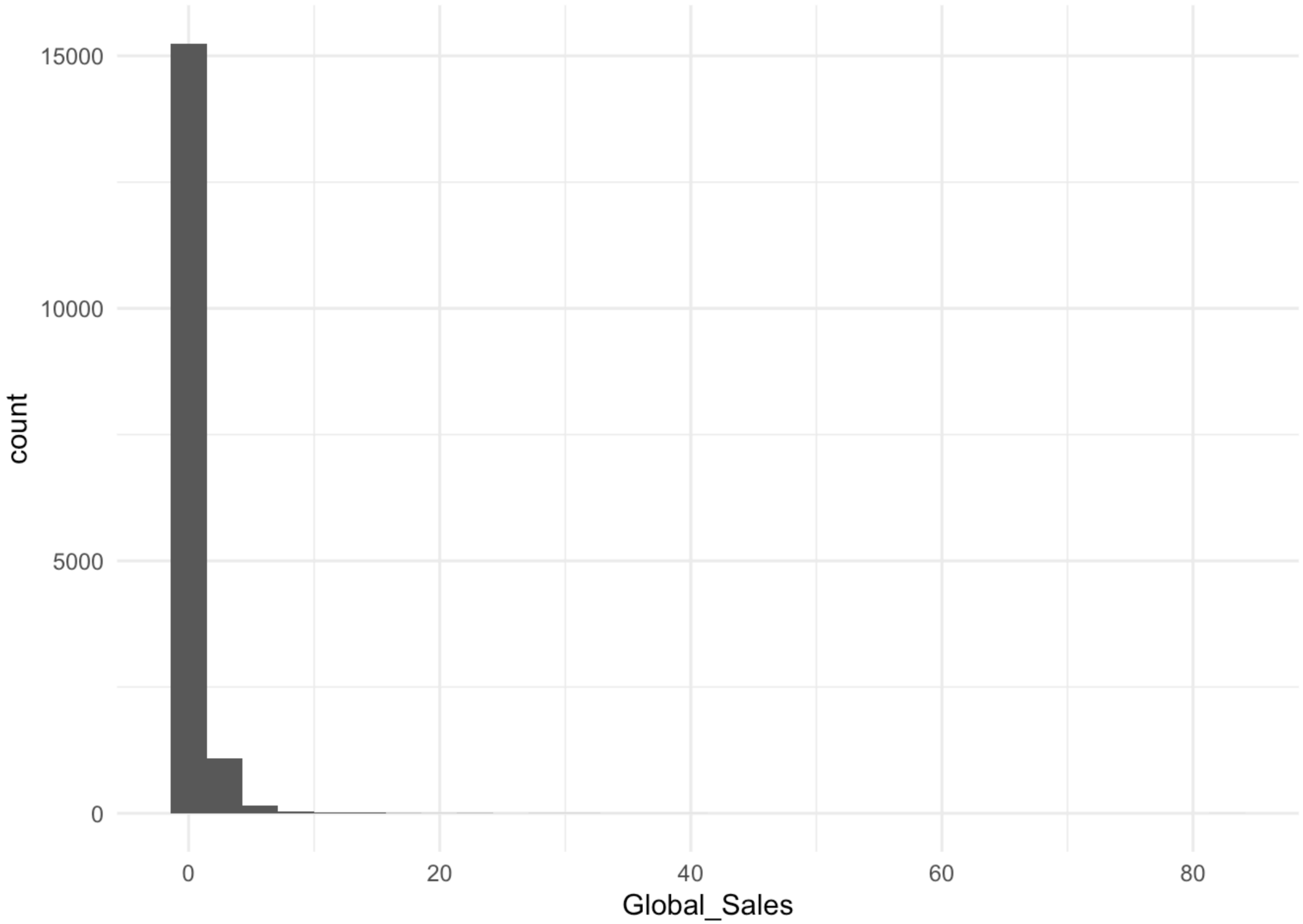
```
data <- read_csv("../data/vgsales.csv")
grouped <- data %>% group_by(Platform) %>% summarise(avg_rank=mean(Rank), avg_sales_m = mean(Global_Sales))
grouped
```

```
## # A tibble: 31 × 3
##    Platform avg_rank avg_sales_m
##    <chr>       <dbl>       <dbl>
##  1 2600        4403.      0.730
##  2 3DO        14373.      0.0333
##  3 3DS         9160.      0.486
##  4 DC          8771.      0.307
##  5 DS          9637.      0.380
##  6 GB          3392.      2.61
##  7 GBA         8682.      0.387
##  8 GC          8664.      0.359
##  9 GEN         7038.      1.05
## 10 GG         13527       0.04
## # ℹ 21 more rows
```

Here we have the average rank and sales per gaming platform.

Now let’s take a look at what this distribution looks like.

```
m <- ggplot(data, aes(Global_Sales)) +
  geom_histogram() + theme_minimal()
m
```

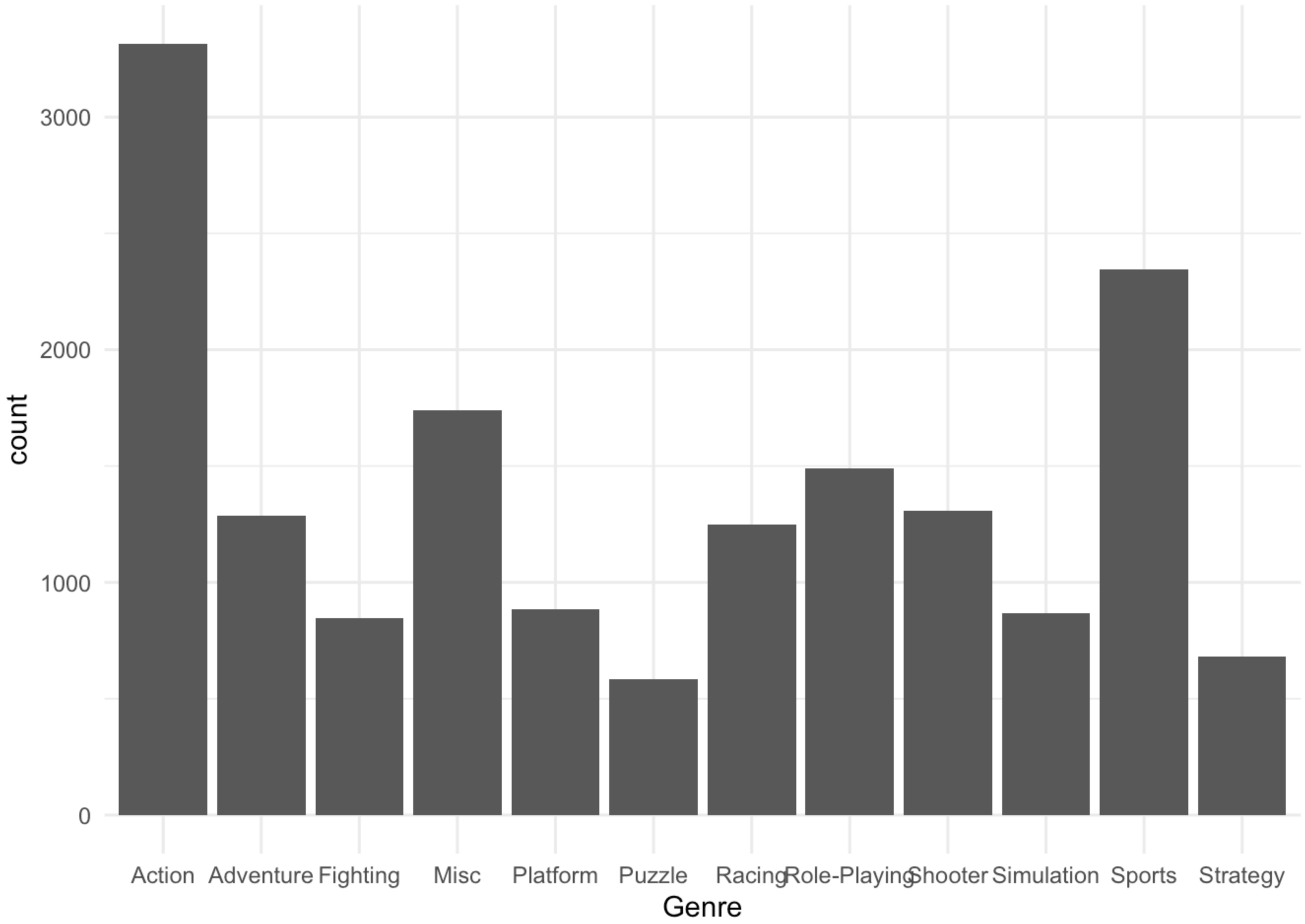


```
max(data$Global_Sales)
```

```
## [1] 82.74
```

As we can see, the global sales volume for most video games tends to be quite low ($0-10m) whereas the max is ~83. However from the histogram, we can see that games with sales volume this high is very rare.

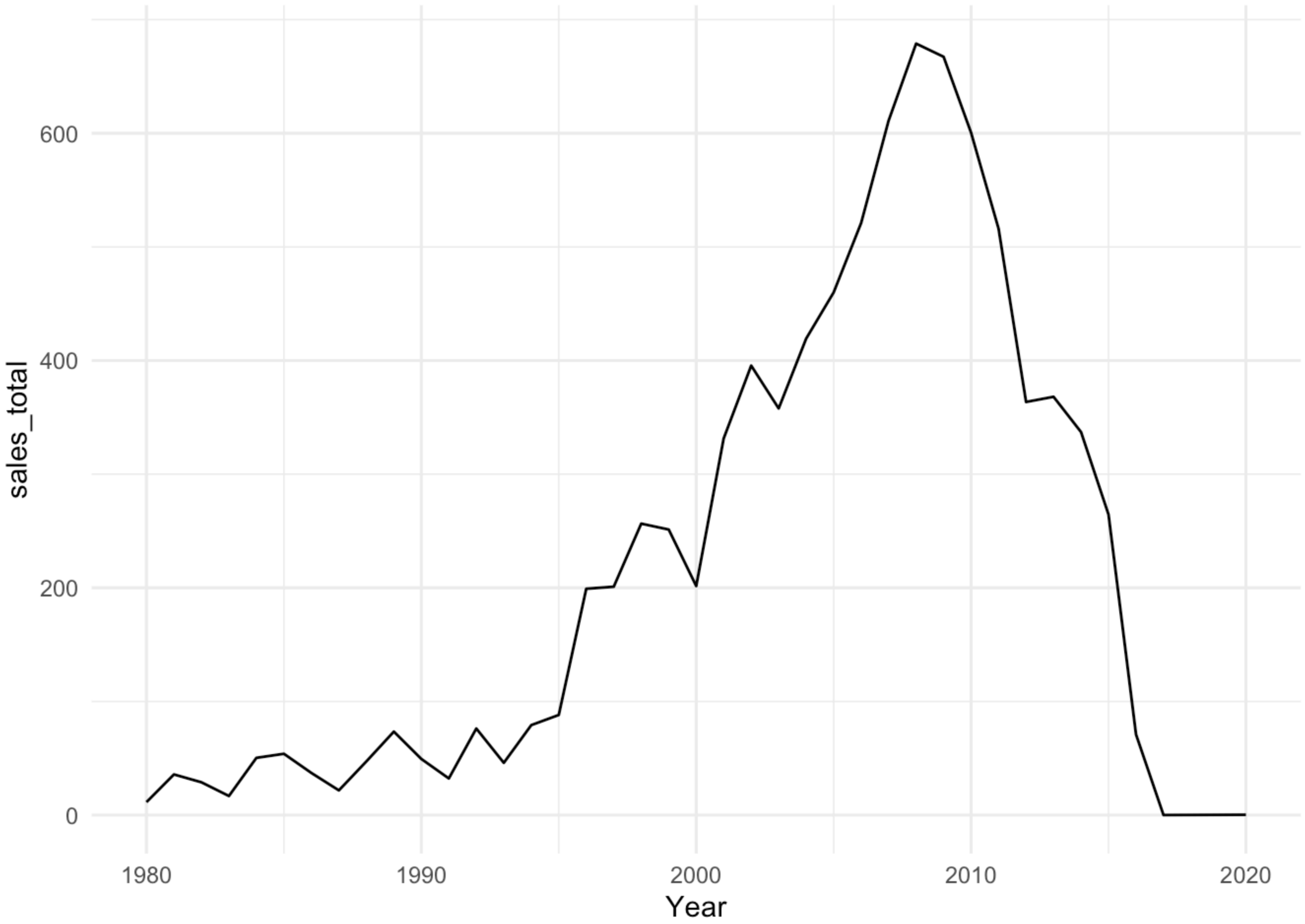
```
genres <- data %>% group_by(Genre) %>% summarize(count=n())
ggplot(genres, aes(Genre, count)) + geom_bar(stat="identity") + theme_minimal()
```



We can also see that Action games are the most popular followed by Sports, with Puzzle games being the least popular.

Finally, let’s take a look at how the sales change over the years.

```
year <- data %>% group_by(Year) %>% summarize(sales_total=sum(Global_Sales))
year$Year <- year$Year %>% as.integer()
ggplot(year, aes(Year, sales_total)) + geom_line() + theme_minimal()
```



From this, we can see that there is a steady decline in the sales volume in the late 2010s. This is inpart due to missing data (i.e no sales information from 2018, 2019) and part of it could be due to covid.

Now that we have done some light exploratory data analysis, let's test the hypothesis.

We formalize it in the following way:

Let X be the distribution of video game sales globally.

$$H_0 : X \sim Poisson(\hat{\lambda}) \text{ versus } H_1 : X \text{ is not Poisson}$$

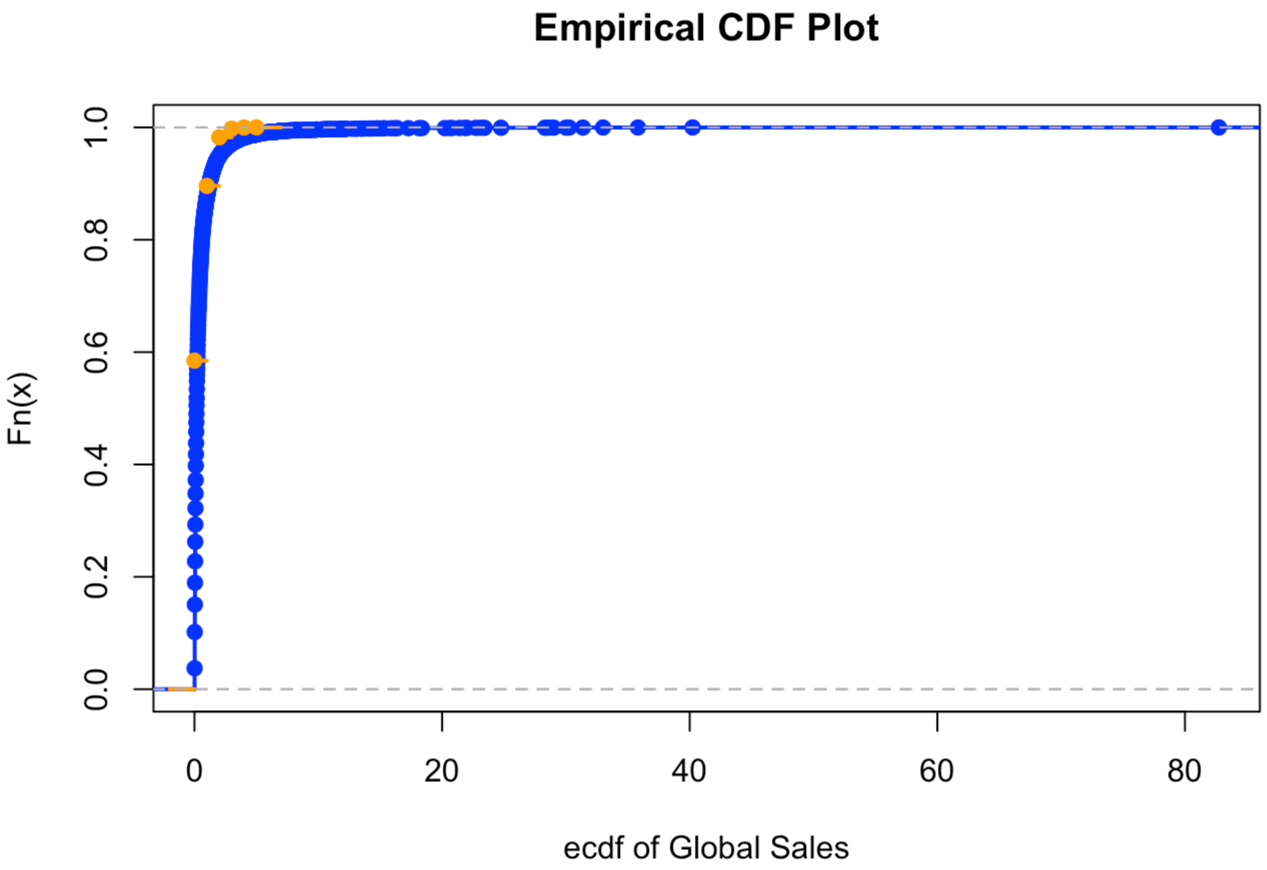
I will set $\hat{lambda}$ to be the MLE so $\hat{\lambda} = \bar{X}$.

I will compare the distribution of the empirical cdf of global sales against a Poisson distribution using the Kolmogorov-Smirnov test. This test rejects for large values of the maximum distance of the ecdf of our data and the cdf for the Poisson distribution with parameter $\hat{\lambda}$.

```
lambda <- data$Global_Sales %>% mean()
x <- data$Global_Sales
y <- rpois(n=length(x), lambda=lambda)
ks.test(x, y)
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  x and y
## D = 0.58477, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
plot(ecdf(x), xlim=range(c(x, y)), verticals=T, xlab="ecdf of Global Sales", col="blue", lwd=2, main="Empirical C
DF Plot")
lines(ecdf(y), col="orange", lwd=2)
```



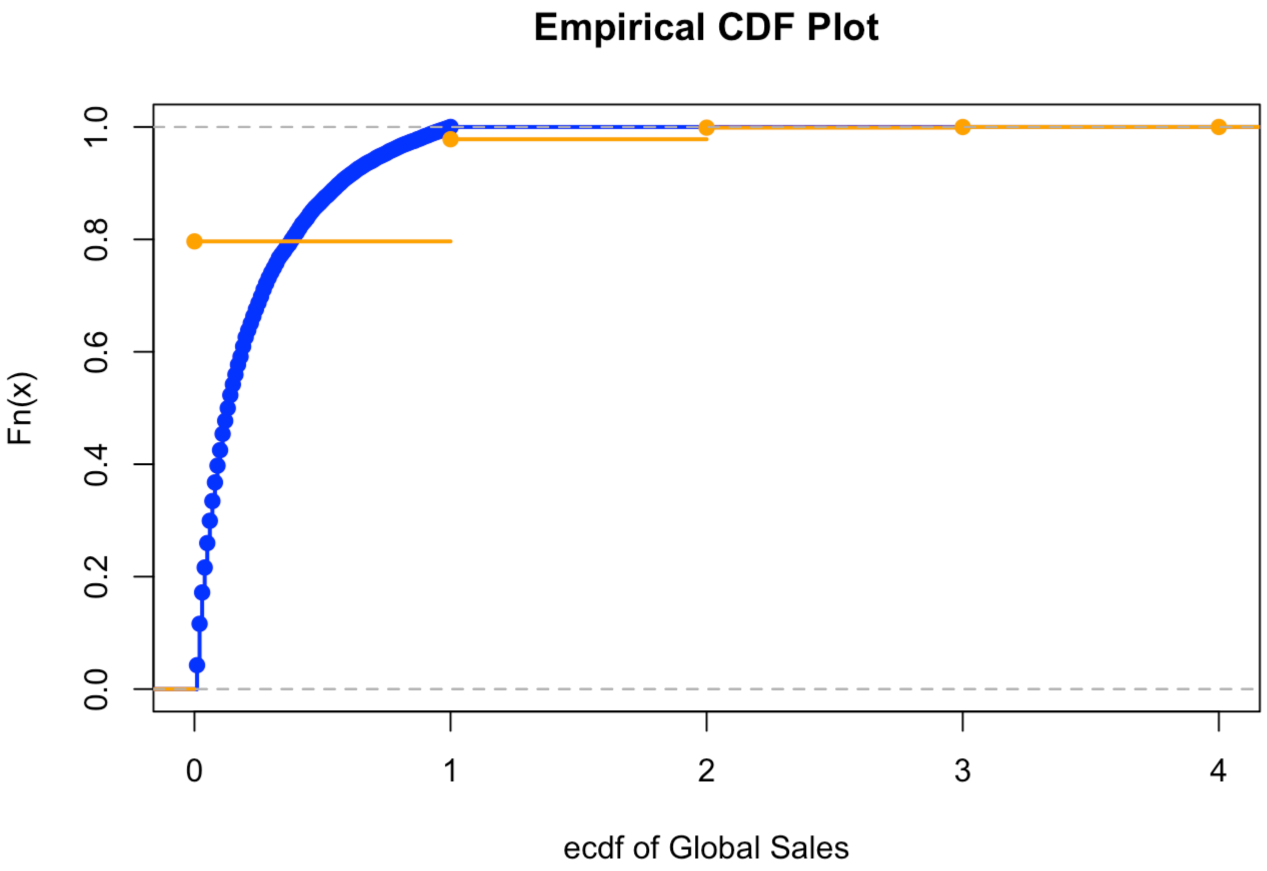
From the picture above, we can see that our distributions look similar, but there is a massive difference at the tail end of the distributions and hence a large distance. This also agrees with our hypothesis test, we reject the null hypothesis, it does not appear that the distribution of video game sales follows a Poisson distribution.

Now we will repeat this process but this time, removing outliers (games with sales > $1m).

```
data_filtered <- data %>% filter(Global_Sales <= 1)
lambda_one <- data_filtered$Global_Sales %>% mean()
x_one <- data_filtered$Global_Sales
y_one <- rpois(n=length(x_one), lambda=lambda_one)
ks.test(x_one, y_one)
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  x_one and y_one
## D = 0.79648, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
plot(ecdf(x_one), xlim=range(c(x_one, y_one)), verticals=T, xlab="ecdf of Global Sales", col="blue", lwd=2, main
="Empirical CDF Plot")
lines(ecdf(y_one), col="orange", lwd=2)
```



Now we see that the gap between the cdfs is due to a difference between the cdf of the Poisson distribution and the ecdf of our data rather than the inverse. Here we also get the same result as the test above and reject the null hypothesis.

Question 2

I define the flipSignTest function as follows:

```
flipSignTest <- function(x, B = 10000) {
  y_star <- mean(x)
  n <- length(x)
  stats = numeric(length = B)
  for (i in 1:B){
    signs <- sample(c(-1, 1), n, replace=TRUE)
    y_eps <- mean(signs * x)
    stats[i] = y_eps
  }
  p = sum(abs(stats) >= abs(y_star)) / 2^n
  return(p)
}
```

Now I will apply this function to the data to compare the video game sales in 2010 between Europe and Japan.

```
data_ten <- data %>% filter(Year=="2010")
europe <- data_ten$EU_Sales
japan <- data_ten$JP_Sales
res_eu <- flipSignTest(europe)
res_jp <- flipSignTest(japan)
c(res_eu, res_jp)
```

```
## [1] 0 0
```

Based on the result of our flipped sign test, it would appear that distribution of video game sales for both Europe and Japan in 2010 do not appear to be congruent with a distribution that is symmetric about 0.

Question 3.

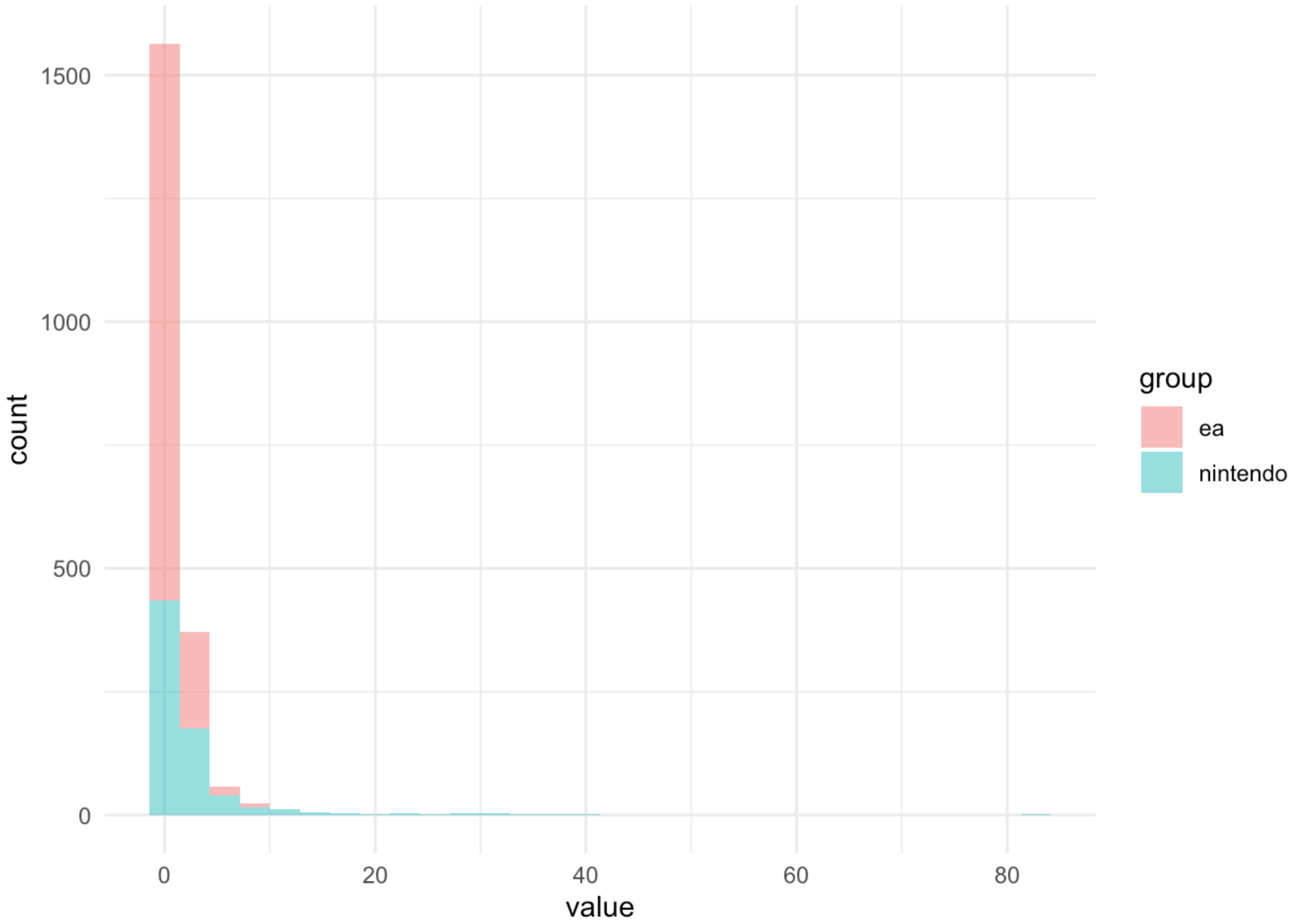
```
data %>% group_by(Publisher) %>% summarise(total_sales = sum(Global_Sales)) %>% arrange(desc(total_sales))
```

```
## # A tibble: 579 × 2
##    Publisher                    total_sales
##    <chr>                              <dbl>
##  1 Nintendo                           1787.
##  2 Electronic Arts                    1110.
##  3 Activision                          727.
##  4 Sony Computer Entertainment         608.
##  5 Ubisoft                             475.
##  6 Take-Two Interactive                400.
##  7 THQ                                 341.
##  8 Konami Digital Entertainment        284.
##  9 Sega                                273.
## 10 Namco Bandai Games                  254.
## # ℹ 569 more rows
```

I am interested to see if the distribution of video game sales between the two top selling publishers are the same or not. From above, we can see that Nintendo takes the top spot with ~$1.8bn in sales with EA following behind at $1.1bn sales. This data is unpaired because the release of a game from Nintendo is independent from the release of a game from EA and thus their sales data are as well.

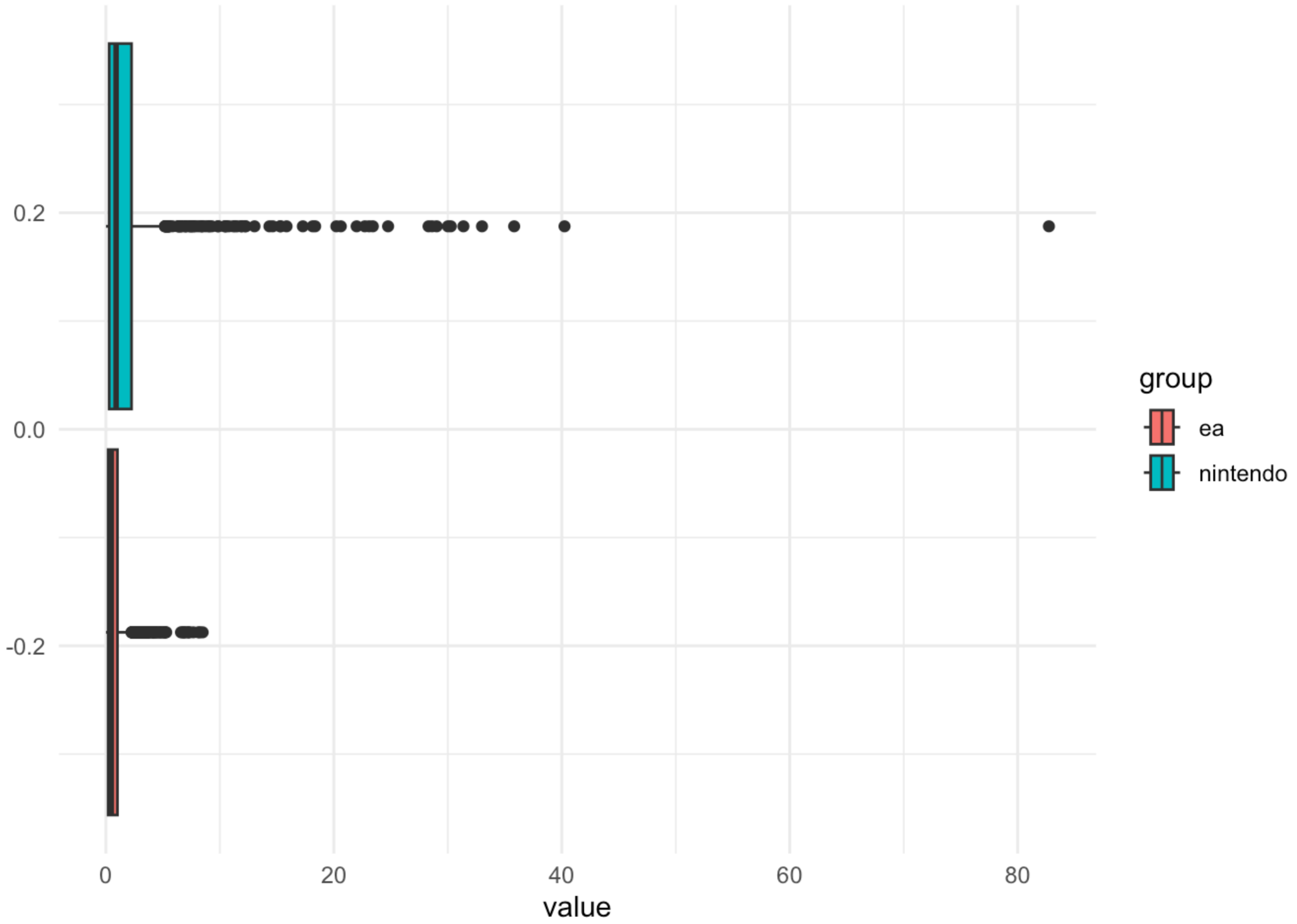
I will conduct further data analyses with these two below.

```
ntndy <- data %>% filter(Publisher == "Nintendo") %>% pull(Global_Sales)
ea <- data %>% filter(Publisher == "Electronic Arts") %>% pull(Global_Sales)
df1 <- data.frame(value = ntndy, group = "nintendo")
df2 <- data.frame(value = ea, group = "ea")
df <- rbind(df1, df2)
ggplot(df, aes(x=value, fill=group)) + geom_histogram(identity="position", alpha=0.5) + theme_minimal()
```



Based on the plot above, the distributions seem to be quite different. Nintendo has more games toward the tail end whereas EA has many games concentrated in the first bin. Let’s take a look at the quantiles of the data.

```
ggplot(df, aes(x=value, fill=group)) + stat_boxplot() + theme_minimal()
```



```
c(summary(ntndy), summary(ea))
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.       Min.    1st Qu.     Median 
##  0.0100000  0.2900000  0.8900000  2.5413371  2.2500000 82.7400000  0.0100000  0.2000000  0.4800000 
##       Mean    3rd Qu.       Max. 
##  0.8218505  1.0100000  8.4900000
```

As we can see from the box plots above, Nintendo has higher median, 75-th percentile, and far more outliers than EA. I have also included the summary for your reference.

Now let's formalize our hypothesis test.

Let F_X be the distribution of Nintendo's sales and F_Y be the distribution of EA's sales.

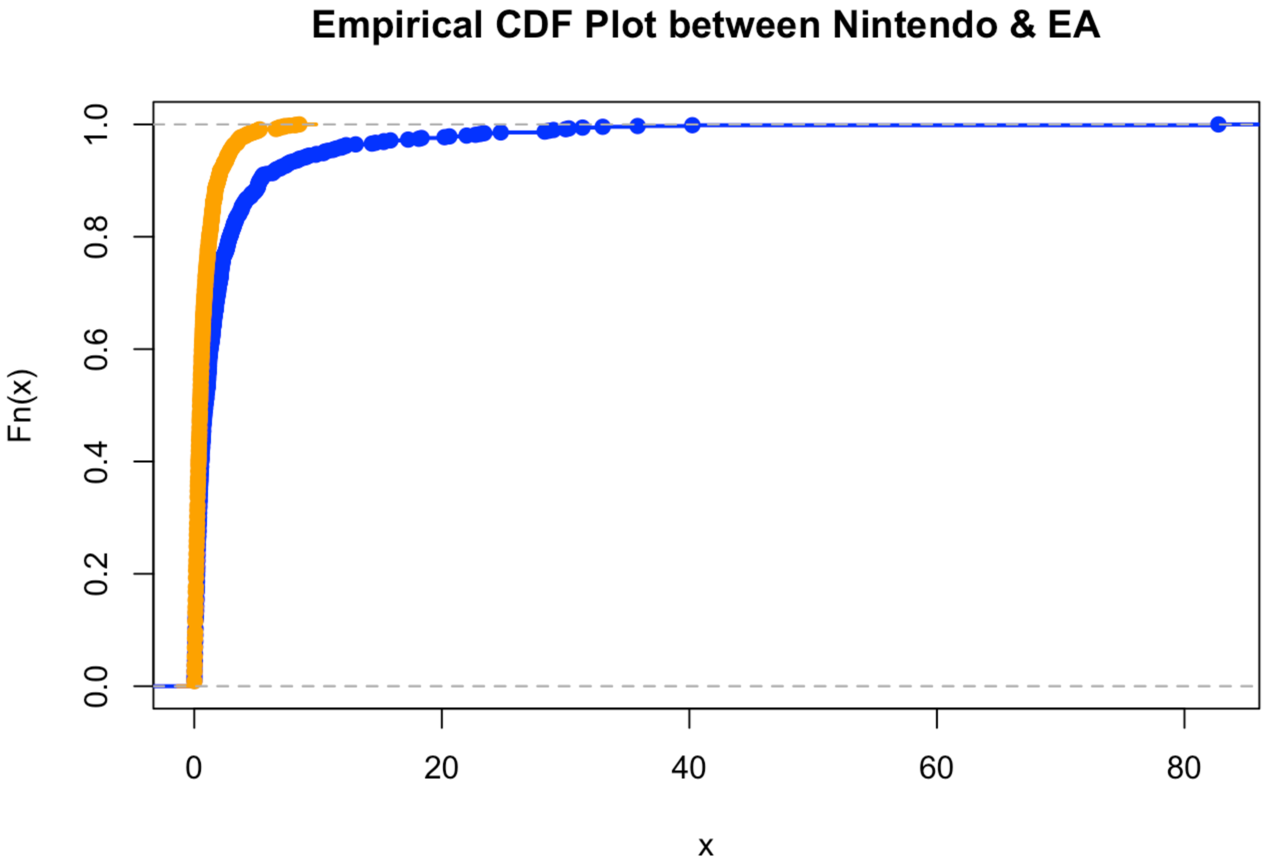
$$H_0 : F_X = F_Y \text{ versus } H_1 : F_X \neq F_Y$$

I will conduct the test below using the Kolmogorov-Smirnov test once again along with an accompanying visual.

```
ks.test(ntndy, ea)
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  ntndy and ea
## D = 0.23797, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
plot(ecdf(ntndy), xlim=range(c(ntndy, ea)), col="blue", lwd=2, main="Empirical CDF Plot between Nintendo & EA")
lines(ecdf(ea), col="orange", lwd=2)
```



As we can see from the plot above, the distributions look quite different. This agrees with the result from our hypothesis test. We reject the null hypothesis, it seems that the distribution of video game sales is different between Nintendo and EA.