Probability and Statistics II Project

Name	*ID*
Zainab Mohamed Abdallah	20221310251
Malak Mahmoud Aref	20221445867
Nureen Ehab Mahmoud	20221465124
Bassant Mohamed	20221376715

Intelligent Systems Department

Introduction:

The main goal of this project to discover what actually happened in this tragic event based on statistics and numeric values. For example: we need to discover statistical effect of lifeboat selection on the survival rate of the passengers based on gender and age. So, with these discoveries we help the largest shipbuilders to take into consideration the necessary numbers of lifeboats that must be on the ship in order to save the largest number of people. Finally, that's the aim for using the data to improve estimates and to achieve more progress and efficiency.

First:

We installed the required packages and load it

```
#install required packages
install.packages("dplyr", dependencies = TRUE)
install.packages("ggplot2", dependencies = TRUE)
install.packages("statsr", dependencies = TRUE)
#load required packages
library(dplyr)
library(ggplot2)
library(statsr)
```

Then we downloaded the dataset and read it using the following code

```
#load the population data
titanic <- read.csv("C:/Users/LENOVO/Downloads/train.csv")
titanic
```

Our data has 891 passengers → females=314, males=577

Q1) We will summarize the statistics of population data but fist, we should remove the missing data to run the code without error (will talk about missing data in the following questions)

```
#summarizes the statistics of the population
titanic_dropna %>%
  summarise(mu = mean(Age), pop_med = median(Age),
  pop_min=min(Age), pop_max=max(Age), sigma = sd(Age),
  pop_q1=quantile(Age,0.25), #firstquartile,25thpercentile
  pop_q3=quantile(Age,0.75)) #thirdquartile,75thpercentile
```

Output:

```
      mu pop_med pop_min pop_max   sigma pop_q1 pop_q3
29.69912      28    0.42      80 14.5265 20.125     38
```

We concluded that the max age = 80 and min = 0.42

Q2) Quantitative data: refers to any information that can be quantified, counted or measured, and given a numeric value.

i) Variables that belong to quantitative: age, fare, number of siblings and spouses, number of children.

Qualitative data: is descriptive in nature, expressed in terms of language.

ii) Variables that belong to qualitative: sex, passenger class, name, ticket, cabin, embarked, survived and passenger id.

Q3)

i) Yes, we can count the missing data using the following code

```
#count the missing data in age variable
sum(is.na(titanic$Age))
```

Output: 177 missing data

ii) Yes, it is normal for the dataset in real world to contain missing data but we can't do operations on dataset that contains missing data, so we have to remove the missing data first.

iii) Estimate mean age of missing females and males

This code gets all the needed information of the missing data

```
#while calculating the average age, we noticed that the Titanic data is incomplete.
#Let's see who these passengers are
Titanic_na <- titanic %>%
  filter(is.na(Age))
```

Example of the Output:

Filter

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583		Q
2	18	1	2	Williams, Mr. Charles Eugene	male	NA	0	0	244373	13.0000		S
3	20	1	3	Masselmani, Mrs. Fatima	female	NA	0	0	2649	7.2250		C
4	27	0	3	Emir, Mr. Farred Chehab	male	NA	0	0	2631	7.2250		C
5	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	NA	0	0	330959	7.8792		Q

Then separate the age column based on gender

```
#Separate the age column of the titanic dataset into two groups based on gender
male = Titanic_na$Age[Titanic_na$Sex == 'male']
female = Titanic_na$Age[Titanic_na$Sex == 'female']
missingdata1 <- data.frame(age_male=male)
missingdata2 <- data.frame(age_female=female)
```

we conclude that number of missing male = 124 and female = 53

Also, when we summarized the age of 714 passengers without the missing data we have known that the maximum age was 80 and the minimum was 0.42 and the mean = 29

Also, from the following questions we will know that the mean of age of males & females of the non-missing data

```
mean(Dataset1age$age_male)
L] 30.72664
mean(Dataset2age$age_female)
L] 27.91571
```

And we noticed earlier that number of missing male = 124 and female = 53

So we conclude that: estimation of average of age of males is in range between 30:31

And estimation of average of age of females is in range between 27:28

We can also know the number of missing females and males if they survived or not by the following code

```
gender_survived_ratio <- Titanic_na %>%
  group_by(Sex, Survived) %>%
  summarise(Count = n()) %>%
  mutate(Percentage = round(Count/sum(Count)*100))
gender_survived_ratio
```

Output:

Sex	Survived	Count	Percentage
female	0	17	32
female	1	36	68
male	0	108	87
male	1	16	13

0 means didn't survived and 1 means survived

Total of passenger who survived = 52 and who didn't survived = 125

We notice that number of survived females greater than number of survived males and that's reasonable as women have priority in the lifeboats than men

iv) Yes, the missing data will affect our statistics because we can't do operations on dataset that contains missing data.

Q4) We will remove the missing data by the following code

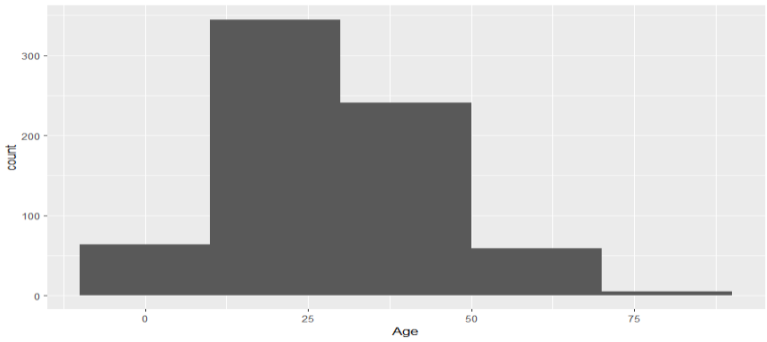
```
#remove the missing data and i save it in variable called titanic_dropna
#so I can keep both the original dataset and also the modified dataset in the working environment
titanic_dropna=na.omit(titanic)
```

Q5) Histogram for the dataset's Age variable.

Code:

```
ggplot(data=titanic_dropna,aes(x=Age))+geom_histogram(binwidth = 20)
```

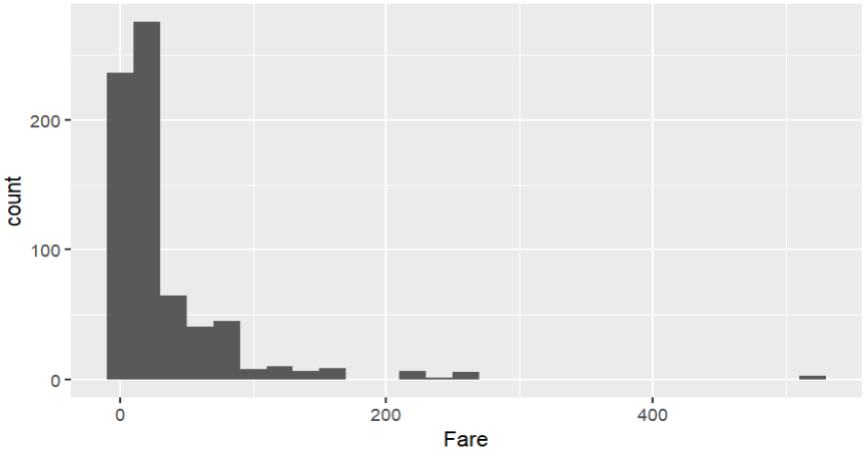
Output:



Q6) The best distribution that may be employed to provide a good fit for the Age and Fare histograms is

normal distribution for Age

right-skewed distribution for Fare



Q7)

mean → 29.69912 standard deviation → 14.5265

```
#Q7
mean(titanic_dropna$Age)
sd(titanic_dropna$Age)
```

Q8) Random sample of size 50

Code:

```
#Q8
sampl<-titanic_dropna%>%
  sample_n(size=50)

sampl%>%
  summarise(mu_sample = mean(Age), sample_med = median(Age),
            sd_sample = sd(Age), sample_iqr = IQR(Age),
            sample_min=min(Age),sample_max=max(Age),
            sample_q1=quantile(Age,0.25),#firstquartile,25thpercentile
            sample_q3=quantile(Age,0.75))#thirdquartile,75thpercentile
```

Output:

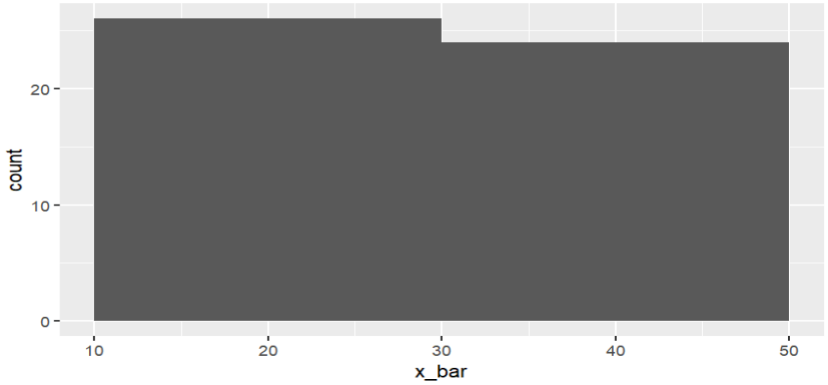
mean_sample → 27.9766 sd_sample → 14.83699

Q9)

Code:

```
#Q9
sample_means50 <- titanic_dropna %>%
  rep_sample_n(size=50, reps=50, replace=TRUE) %>%
  summarise(x_bar = mean(Age))
ggplot(data = sample_means50,aes(x = x_bar)) + geom_histogram(binwidth=20)
mean(sample_means50$x_bar)
sd(sample_means50$x_bar)
```

Output: mean → 30.13374 sd → 2.410372



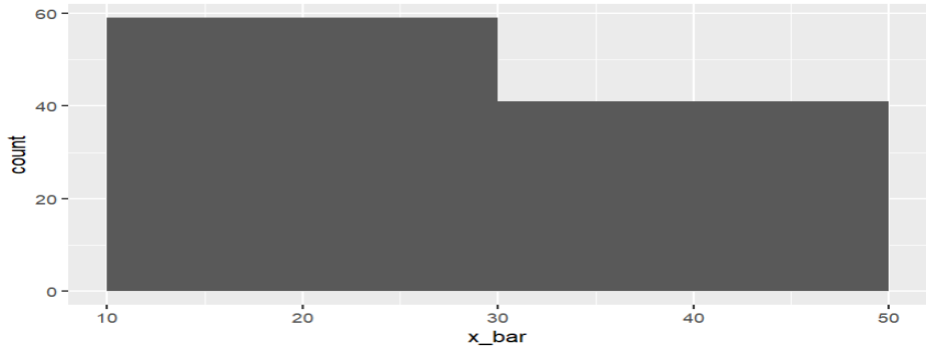
Description of the graph → Normal distribution

Q10)

Code:

```
#Q10
sample_means100 <- titanic_dropna %>%
  rep_sample_n(size=50, reps=100, replace=TRUE) %>%
  summarise(x_bar = mean(Age))
ggplot(data = sample_means100,aes(x = x_bar)) + geom_histogram(binwidth=20)
mean(sample_means100$x_bar)
sd(sample_means100$x_bar)
```

Output: mean → 29.61458 sd → 2.035246



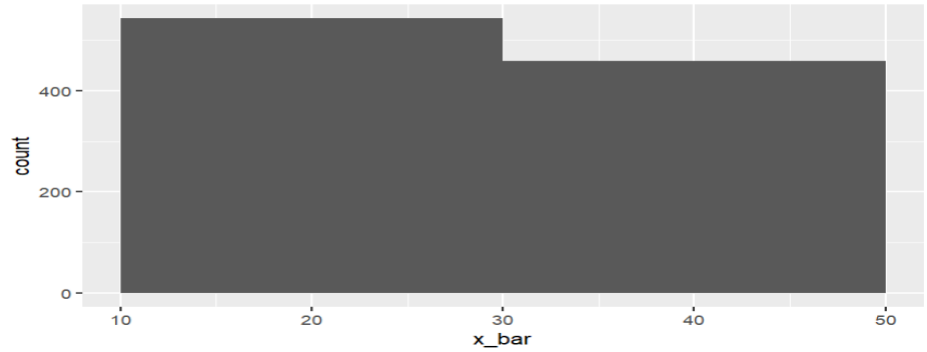
Description of the graph → Normal distribution

Q11)

Code:

```
#Q11
sample_means1000 <- titanic_dropna %>%
  rep_sample_n(size=50, reps=1000, replace=TRUE) %>%
  summarise(x_bar = mean(Age))
ggplot(data = sample_means1000,aes(x = x_bar)) + geom_histogram(binwidth=20)
mean(sample_means1000$x_bar)
sd(sample_means1000$x_bar)
```

Output: mean → 29.75827 sd → 2.080082



Description of the graph → Normal distribution

Q9-Q11 → the histogram is normal distribution and when we reduce the number of binwidth it's being more obvious that it follows normal distribution.

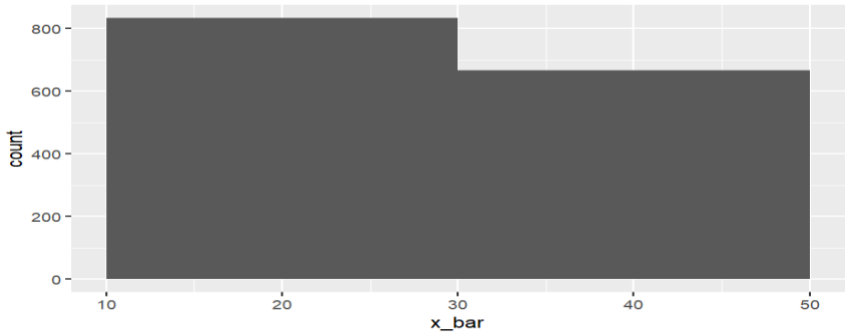
Q12) From Q9-Q11 the mean will be more accurate while increasing the number of samples (repetition) and the standard error has no difference.

Q13)

Code:

```
#Q13
sample_means_s20 <- titanic_dropna %>%
  rep_sample_n(size=20, reps=1500, replace=TRUE) %>%
  summarise(x_bar = mean(Age))
ggplot(data = sample_means_s20,aes(x = x_bar)) + geom_histogram(binwidth=20)
mean(sample_means_s20$x_bar)
sd(sample_means_s20$x_bar)
```

Output: mean → 29.68402 sd → 3.198098



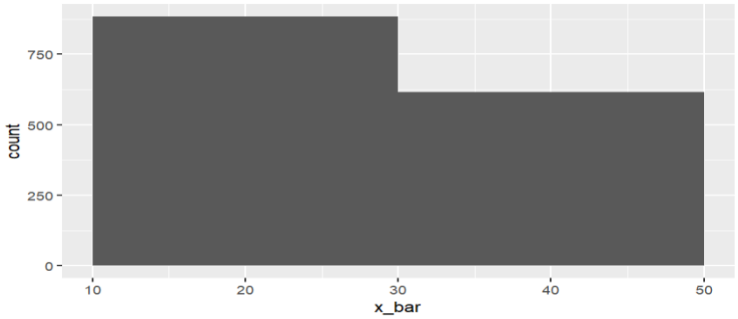
Description of the graph → Normal distribution

Q14)

Code:

```
#Q14
sample_means_s100 <- titanic_dropna %>%
  rep_sample_n(size=100, reps=1500, replace=TRUE) %>%
  summarise(x_bar = mean(Age))
ggplot(data = sample_means_s100,aes(x = x_bar)) + geom_histogram(binwidth=20)
mean(sample_means_s100$x_bar)
sd(sample_means_s100$x_bar)
```

Output: mean → 29.65856 sd → 1.467422



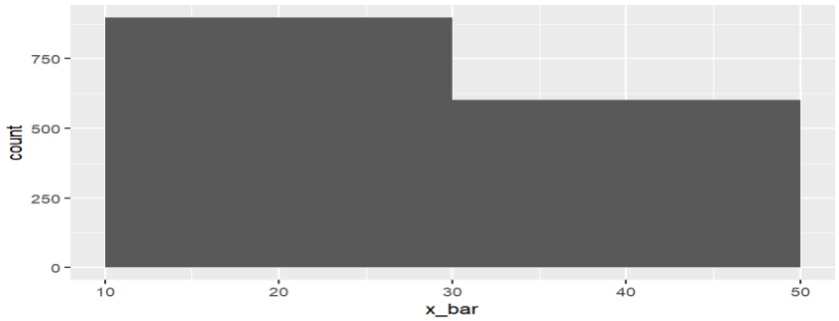
Description of the graph → Normal distribution

Q15)

Code:

```
#Q15
sample_means_s200 <- titanic_dropna %>%
  rep_sample_n(size=200, reps=1500, replace=TRUE) %>%
  summarise(x_bar = mean(Age))
ggplot(data = sample_means_s200,aes(x = x_bar)) + geom_histogram(binwidth=20)
mean(sample_means_s200$x_bar)
sd(sample_means_s200$x_bar)
```

Output: mean → 29.71706 sd → 1.035535



Description of the graph → Normal distribution

Q16)

When sample size increases the standard error (variability) decreases as for the centeral limit theorem: from any population with mean μ and variance sigma^2 then for large enough n≥30, xbar follows (μ,sigma^2/n)

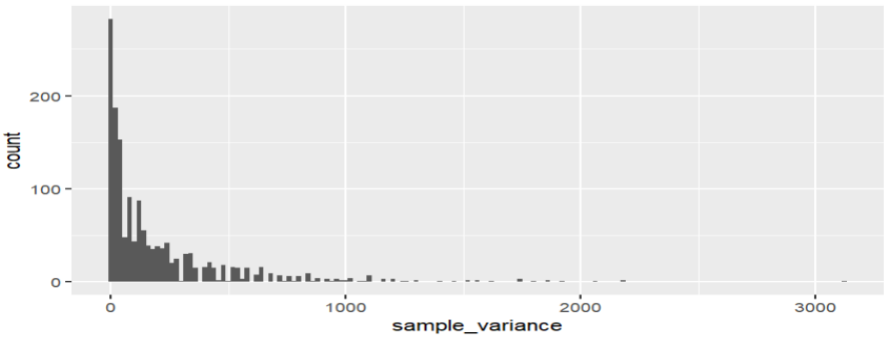
but in Q13 → n = 20 so the standard error was very large.

Q17) The shape of histogram doesn't follow normal distribution it follows chi square degree of freedom = 1

Code:

```
sample_U1500 <- titanic_dropna %>%
rep_sample_n(size=2, reps=1500, replace=TRUE) %>%
summarise(sample_variance = var(Age))
ggplot(data = sample_U1500,aes(x = sample_variance)) + geom_histogram(binwidth=20)
```

Output:



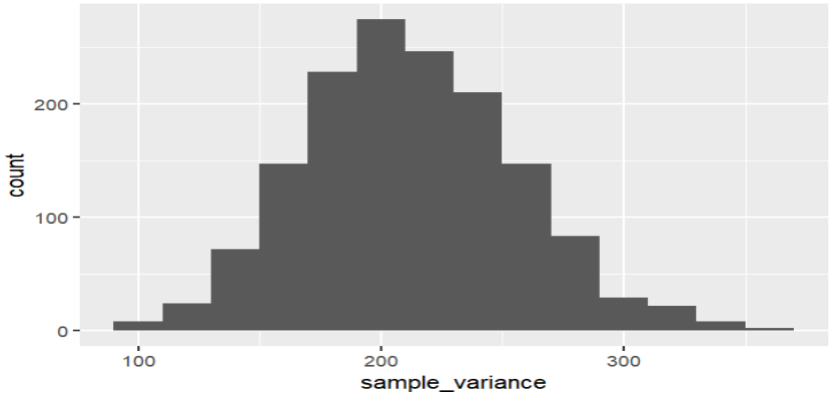
Description of the graph → Chi distribution of dof = 1

Q18) The shape of histogram doesn't follow normal distribution it follows chi square degree of freedom 49

Code:

```
#the histogram will be bigger because of the dof
sample_U1500 <- titanic_dropna %>%
rep_sample_n(size=50, reps=1500, replace=TRUE) %>%
summarise(sample_variance = var(Age))
ggplot(data = sample_U1500,aes(x = sample_variance)) + geom_histogram(binwidth=20)
```

Output:



i) The curve of chi-square distribution is not symmetric and as degree of freedom increase the graph will slightly shifted to the right.

ii) Estimate the population variance: $Var(s^2)=2\sigma^4/(n-1)$

iii) if I have sampling distribution of means can I get the population variance? Yes, the variance of the sampling distribution of the mean is the population variance divided by N.

Q19) n=50

Method of Moments Code:

```
n=50
x=rnorm(n, mean = mean(titanic_dropna$Age))
mean_est=sum(x)/n
mean_est
bias(29.69912,29.53581)
```

Output:

```
> mean_est
[1] 29.53776
> bias(29.69912,29.53776)
[1] 0.16136
```

Maximum Likelihood Code:

```
sample = rnorm(50,mean=29.69912,sd=14.5265)
NLL = function(pars, data) {
  mu = pars[1]
  sigma = pars[2]
  -sum(dnorm(x = data, mean = mu, sd = sigma, log = TRUE))
}
mle = optim(par = c(mu = 0.2, sigma = 1.5), fn = NLL, data = sample, control = list(parscale = c(mu = 0.2, sigma =1.5)))
mle$par
bias(29.69912,29.07958)
```

Output:

```
> mle$par
       mu     sigma
29.07958 13.71102
> bias(29.69912,29.07958)
[1] 0.61954
```

Q20) n=200

Method of Moments Code:

```
n=200
x=rnorm(n, mean =mean(titanic_dropna$Age))
mean_est=sum(x)/n
mean_est
bias(29.69912,29.64748)
```

Output:

```
> mean_est
[1] 29.64748
> bias(29.69912,29.64748)
[1] 0.05164
```

Maximum Likelihood Code:

```
sample = rnorm(200,mean=29.69912,sd=14.5265)
NLL = function(pars, data) {
  mu = pars[1]
  sigma = pars[2]
  -sum(dnorm(x = data, mean = mu, sd = sigma, log = TRUE))
}
mle = optim(par = c(mu = 0.2, sigma = 1.5), fn = NLL, data = sample, control = list(parscale = c(mu = 0.2, sigma =1.5)))
mle$par
bias(29.69912,29.40505)
```

Output:

```
> mle$par
       mu     sigma
29.40505 14.29585
> bias(29.69912,29.40505)
[1] 0.29407
```

Conclusion: The bias decreases as sample size increases and from Q19-Q20 we concluded that the MME is better than MLE because its produce smaller value of bias.

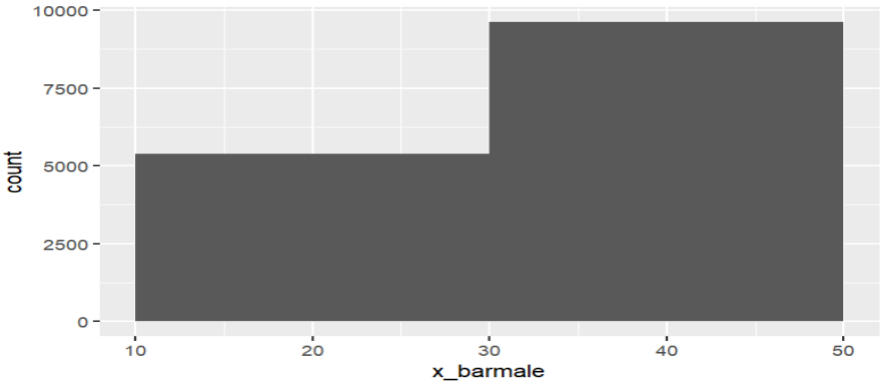
Q21) Code:

```
#Separate the age column of the titanic dataset into two groups based on gender
male = titanic$Age[titanic$Sex == 'male']
male = na.omit(male)
female = titanic$Age[titanic$Sex == 'female']
female = na.omit(female)
Dataset1age <- data.frame(age_male=male)
Dataset2age <- data.frame(age_female=female)

#sampling distribution for age_male
samplediff_means15000 <- Dataset1age %>%
  rep_sample_n(size=50, reps=15000, replace=TRUE) %>%
  summarise(x_barmale = mean(age_male))
ggplot(data = samplediff_means15000,aes(x = x_barmale)) + geom_histogram(binwidth=20)
mean(samplediff_means15000$x_barmale)

#sampling distribution for age_female
samplediff_means15000 <- Dataset2age %>%
  rep_sample_n(size=50, reps=15000, replace=TRUE) %>%
  summarise(x_barfemale = mean(age_female))
ggplot(data = samplediff_means15000,aes(x = x_barfemale)) + geom_histogram(binwidth=20)
mean(samplediff_means15000$x_barfemale)
```

```
> mean(Dataset1age$age_male)
[1] 30.72664
> mean(Dataset2age$age_female)
[1] 27.91571
```

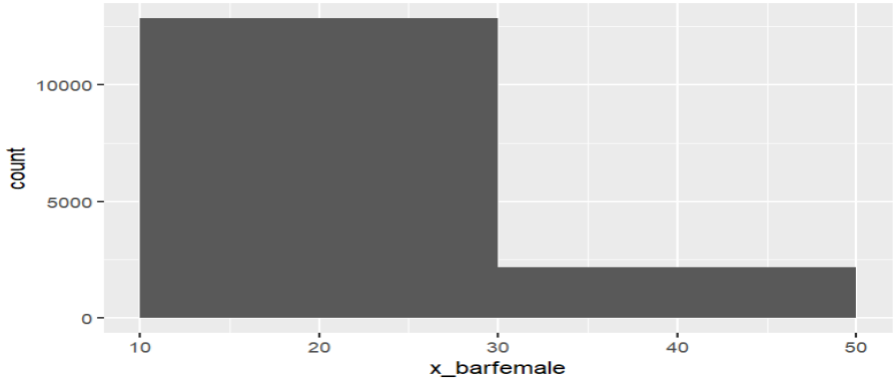
Output:

Sampling distribution of age_male Mean = 30.70902



Description of the graph → Normal distribution

Sampling distribution of age_female Mean = 27.93487



Description of the graph → Normal distribution

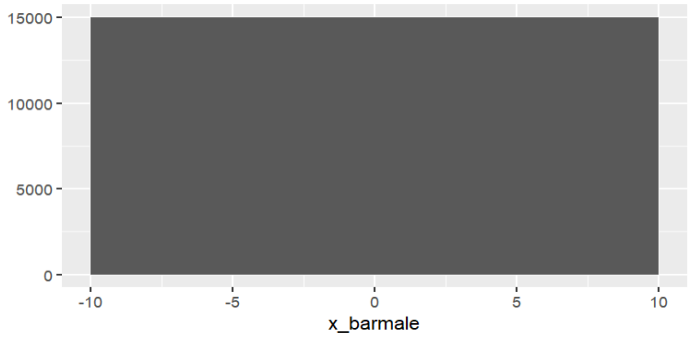
Do you believe there is a significant difference in the average ages of men and women in titanic? Yes, there is a difference because number of males = 453 which is bigger than number of females = 261 also, from the graphs we concluded that the number of older people in males are larger than females in range 30-50 and vice versa in range 10-30.

Q22) Code:

```
#Q22
#Separate the survived column of the titanic dataset into two groups based on gender
survivedmale = titanic_dropna$Survived[titanic_dropna$Sex == 'male']
survivedfemale = titanic_dropna$Survived[titanic_dropna$Sex == 'female']
Dataset1survived <- data.frame(survived_male=survivedmale)
Dataset2survived <- data.frame(survived_female=survivedfemale)

#sampling distribution for survived_male with mean of x_bar of males
samplediff_Survived15000 <- Dataset1survived %>%
  rep_sample_n(size=50, reps=15000, replace=TRUE) %>%
  summarise(x_barmale = mean(survived_male))
ggplot(data = samplediff_Survived15000,aes(x = x_barmale)) + geom_histogram(binwidth=20)
mean(samplediff_Survived15000$x_barmale)

#sampling distribution for survived_female with mean of x_bar of females
samplediff_Survived15000 <- Dataset2survived %>%
  rep_sample_n(size=50, reps=15000, replace=TRUE) %>%
  summarise(x_barfemale = mean(survived_female))
ggplot(data = samplediff_Survived15000,aes(x = x_barfemale)) + geom_histogram(binwidth=20)
mean(samplediff_Survived15000$x_barfemale)
```

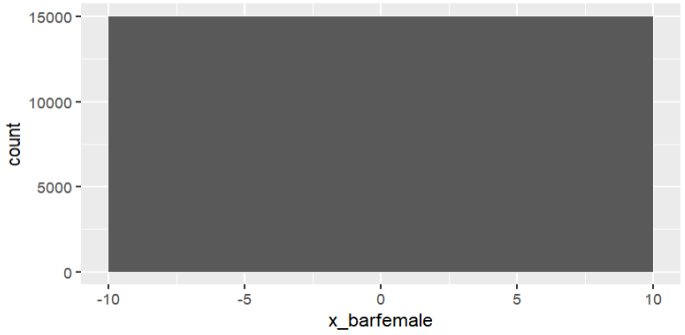
Output:

Sampling distribution of survived_male with mean of xbar of males



Description of the graph → Normal distribution Mean = 0.2059933

Sampling distribution of survived_female with mean of xbar of females

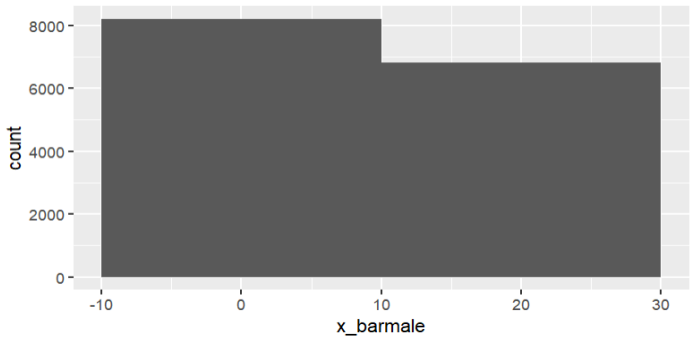


Description of the graph → Normal distribution Mean = 0.7556933

We conclude also from the mean we see that the mean of females more closer to 1 means that the females who survived more than the males who survived

Same code but we change the x_bar = sum not mean

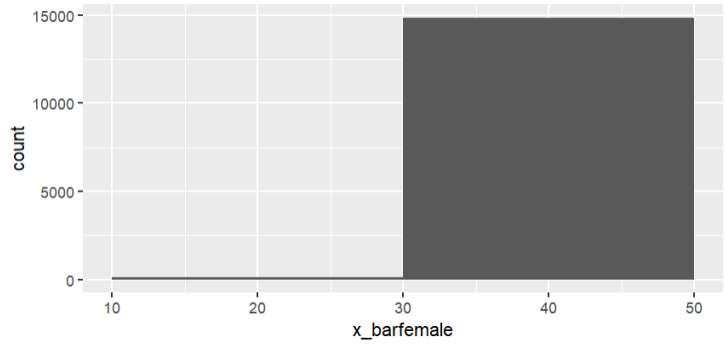
Sampling distribution of survived_male with sum of xbar of males



Mean= 10.26893

Description of the graph → Normal distribution

Sampling distribution of survived_female with sum of xbar of females



Description of the graph → Normal distribution

Mean= 37.74373

Based on this sampling distribution, do you think there is bias in the rescue process between males and females? Yes, because the priority is to rescue the females, children and the older people more than the males that's why we can see that the mean of females bigger than the mean of males.

Q23) Code:

n<30 and age is normal distribution so we used the z

```
#ave +- z*se
mean <- 29.69912
sd <- 14.5265
n <- 10
error <- qnorm(0.05, lower.tail = FALSE)*sd/sqrt(n)
mean - error
mean + error
```

Output: $22.14318 \leq \mu \leq 37.25506$

Q24) Code:

n>30 and age is normal distribution so we used the z

```
#ave +- z*se
mean <- 29.69912
sd <- 14.5265
n <- 50
error <- qnorm(0.05, lower.tail = FALSE)*sd/sqrt(n)
mean - error
mean + error
```

Output: $26.32 \leq \mu \leq 33.07824$

We observed that from Q23-Q24 that the range of the interval decreased this is due to that we increased the number of size.

Q25) Code:

```
sampltimes <- titanic_dropna%>%
  sample_n(size=200)
sampltimes %>%
  summarise(mu_sample = mean(5*Age), var_sample = var(5*Age))
```

Output:

```
 mu_sample var_sample
   150.173    5143.445
```

When we multiply the mean and variance by a constant the → const*E(age) and const^2*V(age)

Q26) Code:

```
sampladd <- titanic_dropna%>%
  sample_n(size=200)
sampladd %>%
  summarise(mu_sample = mean(5+Age), var_sample = var(5+Age))
```

Output:

```
mu_sample var_sample
 34.51125    199.3091
```

When we add the mean and variance by a constant the → const+E(age) and V(age)

Q27) Kernel distribution:

A kernel distribution is a nonparametric representation of the probability density function (pdf) of a random variable.

You can use a kernel distribution when a parametric distribution cannot properly describe the data, or when you want to avoid making assumptions about the distribution of the data.

A kernel distribution is defined by a smoothing function and a bandwidth value, which control the smoothness of the resulting density curve.

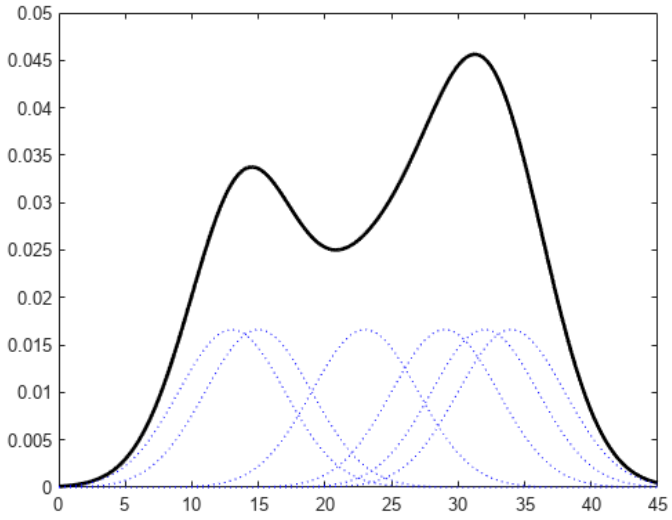
The kernel density estimator is the estimated pdf of a random variable. For any real values of x, the kernel density estimator's formula is given by:

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

The kernel distribution builds a function to represent the probability distribution using the sample data.

A kernel distribution sums the component smoothing functions for each data value to produce a smooth, continuous probability curve.

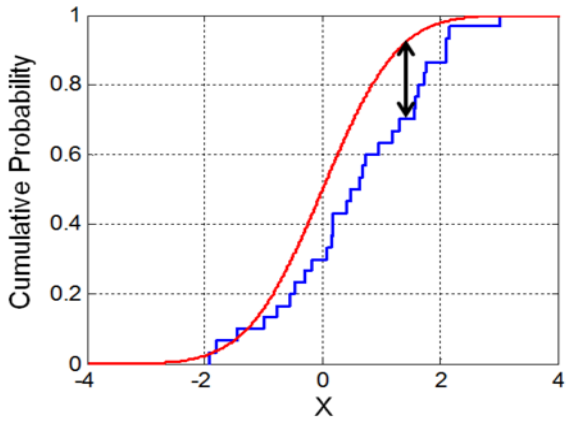
Graph:



Q28) KS-Test:

The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples. The null distribution of this statistic is calculated under the null hypothesis that the sample is drawn from the reference distribution (in the one-sample case) or that the samples are drawn from the same distribution (in the two-sample case). In the one-sample case, the distribution considered under the null hypothesis may be continuous, purely discrete or mixed. In the two-sample case, the distribution considered under the null hypothesis is a continuous distribution but is otherwise unrestricted. However, the two sample test can also be performed under more general conditions that allow for discontinuity, heterogeneity and dependence across samples.

The two-sample K–S test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.



How to run the test by hand

The hypotheses for the test are:

Null Hypothesis (H0): the data comes from the specified distribution.

Alternative Hypothesis (H1): at least one value does not match the specified distribution.
H0: P = P0, Ha: P ≠ P0.
Where P is the distribution of your sample (i.e. the EDF) and P0 is a specified distribution.

General Steps to run the test are:

1. Create an EDF for your sample data
2. Specify a parent distribution (one that you want to compare your EDF to).
3. Graph the two distributions together.
4. Measure the greatest vertical distance between the two graphs.
5. Calculate the test statistic.
6. Find the critical value in the KS table.
7. Compare to the critical value.

Calculating the Test Statistic

The K-S test statistic measures the largest distance between the EDF $F_{data}(x)$ and the theoretical function $F_0(x)$, measured in a vertical direction (Kolmogorov as cited in Stephens 1992).

The test statistic is given by:

$$D = \sup_x |F_0(x) - F_{data}(x)|$$

Where (for a two-tailed test):

$F_0(x)$ = the cdf of the hypothesized distribution,

$F_{data}(x)$ = the empirical distribution function of your observed data.

For a one-tailed test, omit the absolute values from the formula.

If D is greater than the critical value, the null hypothesis is rejected. Critical values for D are found in the KS table P-Value Table.