

Christy 24

This report conducts statistic analysis of bank customers, based on a database including data of 200 customers' annual gross income and education level.

Task 1

First part would be some distribution plots and descriptive statistics of the data, and fitted distribution model.

Load data and libraries we needed:

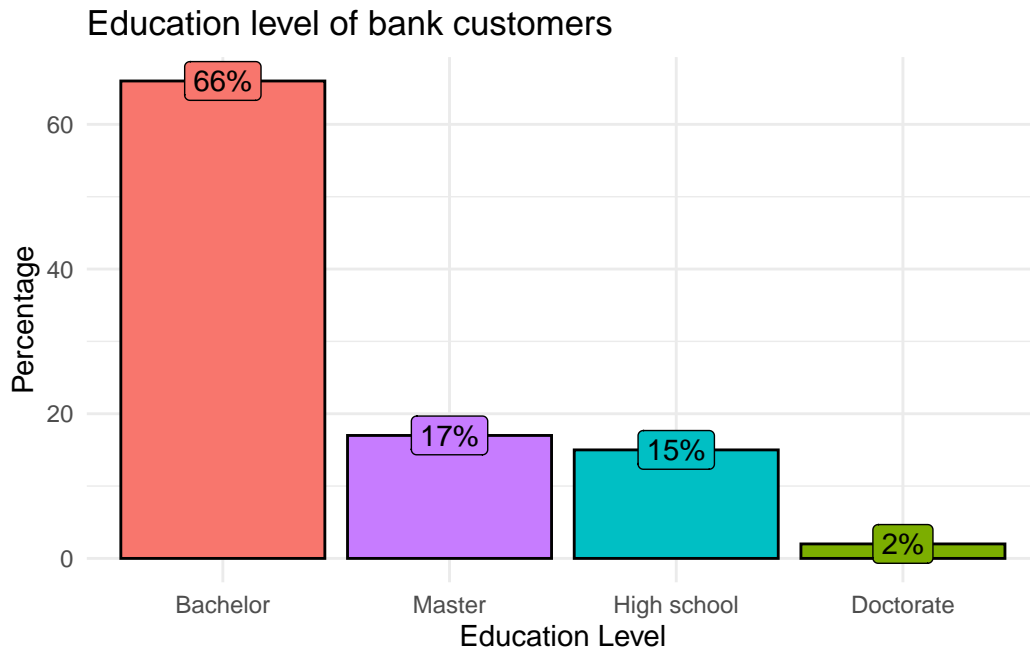
```
# Load necessary libraries
library(tidyverse)
library(kableExtra)
library(fitdistrplus)

# Load the dataset
data <- read_csv("data/banksurvey.csv")
```

The first plot is about the education level of bank customers.

```
# Descriptive bar plot
data %>%
  group_by(education) %>%
  summarise(perc = n()/nrow(data) * 100) %>%
  as.data.frame() %>%
  ggplot(aes(x = fct_reorder(education, -perc),
               y = perc,
               fill = education)) +
  geom_col(color = "black") +
  theme_minimal() +
```

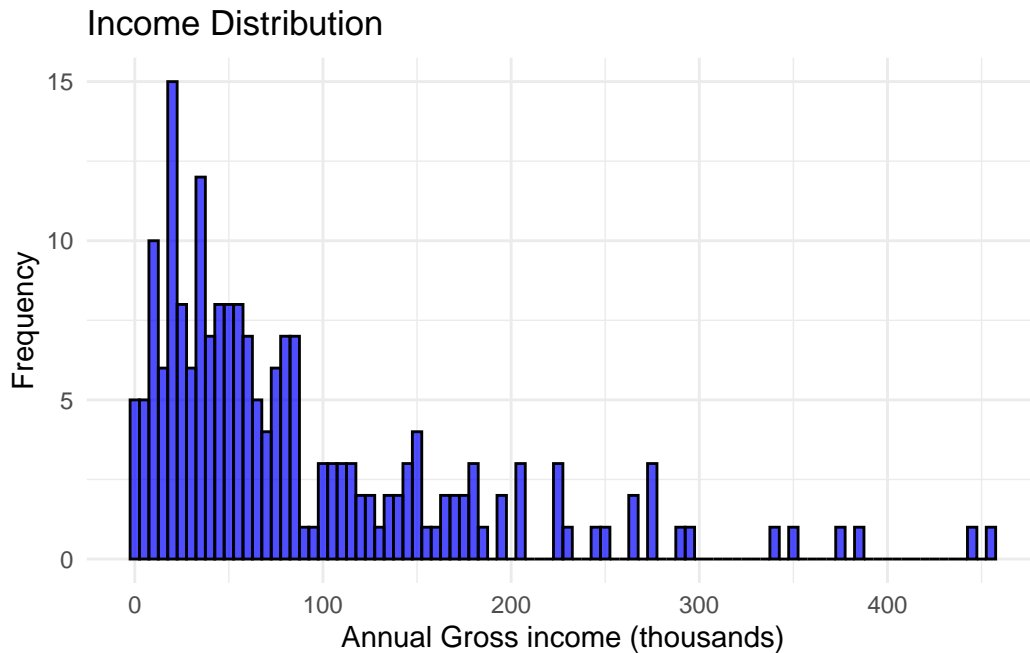
```
labs(x = "Education Level",
     y = "Percentage",
     title = "Education level of bank customers") +
geom_label(aes(label = paste0(perc, "%"),
                          vjust = 0.5)) +
guides(fill = "none")
```



The bar plot shows that most of the surveyed bank customers achieved a highest education level of bachelor's degree (66%), followed by master's and high school with similar percentage, and 2% hold a doctorate degree.

The second part is visualizing the distribution of customer incomes with a histogram, which gives us insight into how incomes are spread across our customer sample.

```
# Histogram of incomes
ggplot(data, aes(x = income)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Income Distribution", x = "Annual Gross income (thousands)", y = "Frequency")
  theme_minimal()
```



The histogram reveals a right-skewed distribution, with most customers having incomes clustered in the lower range and a tail extending toward higher income levels.

We calculate the summary statistics to get more detailed insights into the distribution of incomes.

```
# Descriptive statistics
summary_stats <- data |>
  summarize(
    Mean = round(mean(income), 1),
    Median = median(income),
    SD = round(sd(income), 1),
    Min = min(income),
    Max = max(income)
  )

kable(summary_stats,
  caption = "Statistics of annual gross income of bank customers (thousands)")
```

Table 1: Statistics of annual gross income of bank customers (thousands)

Mean	Median	SD	Min	Max
89.2	58.7	87.6	0.1	453.9

The average income is \$89,200, but the median income of \$58,700 provides a more accurate representation of the typical customer due to the skewed distribution. The standard deviation of \$87,600 reflects a significant variation in incomes.

The next part will be exploring fitting models, including normal, exponential and gamma distribution.

```
# Fit normal distribution
fit_norm <- fitdist(data$income, "norm")
# Fit exponential distribution
fit_exp <- fitdist(data$income, "exp")
# Fit gamma distribution
fit_gamma <- fitdist(data$income, "gamma")

# Normal distribution summary
tibble(Parameter = c("Mean", "Standard Deviation"),
        Estimate = fit_norm$estimate,
        'Standard Error' = fit_norm$sd) %>%
  kable(caption = "Maximum likelihood estimates of fitted normal distribution model")
```

Table 2: Maximum likelihood estimates of fitted normal distribution model

Parameter	Estimate	Standard Error
Mean	89.2250	6.176770
Standard Deviation	87.3526	4.367631

```
# Exponential distribution summary
tibble(Parameter = "Rate",
        Estimate = fit_exp$estimate,
        'Standard Error' = fit_exp$sd) %>%
  kable(caption = "Maximum likelihood estimates of fitted exponential distribution model")
```

Table 3: Maximum likelihood estimates of fitted exponential distribution model

Parameter	Estimate	Standard Error
Rate	0.0112076	0.0007861

```
# Gamma distribution summary
tibble(Parameter = c("Shape", "Rate"),
        Estimate = fit_gamma$estimate,
        'Standard Error' = fit_gamma$sd) %>%
  kable(caption = "Maximum likelihood estimates of fitted gamma distribution model")
```

Table 4: Maximum likelihood estimates of fitted gamma distribution model

Parameter	Estimate	Standard Error
Shape	1.083319	0.0953152
Rate	0.012132	0.0013355

Above tables showed the maximum likelihood estimates of the 3 fitted models.

```
# QQ plots for the normal, exponential, and gamma distributions
par(mfrow = c(2, 2)) # Set up plot window for multiple plots

# QQ plot for normal distribution
qqnorm(data$income)
qqline(data$income,col = "green")
title("Normal QQ Plot")

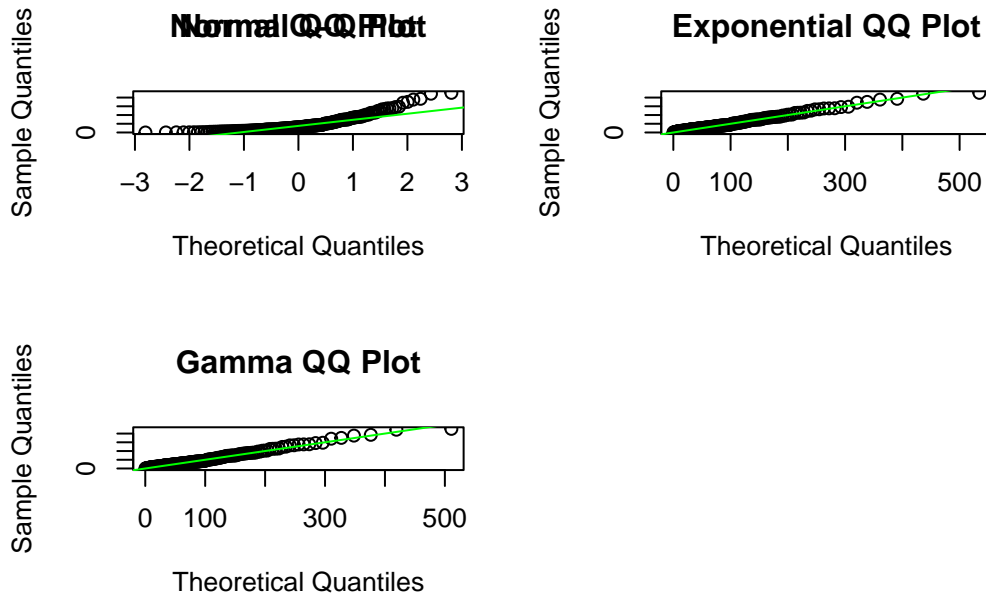
# QQ plot for exponential distribution
qqplot(qexp(ppoints(length(data$income))),
       rate = fit_exp$estimate),
       sort(data$income),
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles")
abline(0, 1,col = "green")
title("Exponential QQ Plot")

# QQ plot for gamma distribution
qqplot(qgamma(ppoints(length(data$income))),
       shape = fit_gamma$estimate[1],
       rate = fit_gamma$estimate[2]),
```

```

    sort(data$income),
    xlab = "Theoretical Quantiles",
    ylab = "Sample Quantiles")
abline(0, 1,col = "green")
title("Gamma QQ Plot")

```



Seems like both exponential and gamma distribution models fit the data better than normal distribution. In particular, exponential distribution is slightly better than gamma.

Task 4

This section will be focusing on the relationship between customer education level and income.

```

# Recode education levels into two groups: "University Degree" vs "No University Degree"
data <- data |>
  mutate(education_group = ifelse(education %in% c("Bachelor", "Master", "Doctorate"),
                                   "University Degree", "No University Degree"))

# Group summary of income by education group
income_education_summary <- data |>

```

```
group_by(education_group) |>
  summarise(mean_income = round(mean(income), 1),
            sd_income = round(sd(income), 1),
            n = n())

kable(income_education_summary,
      caption = "Income comparison of university degree holders")
```

Table 5: Income comparison of university degree holders

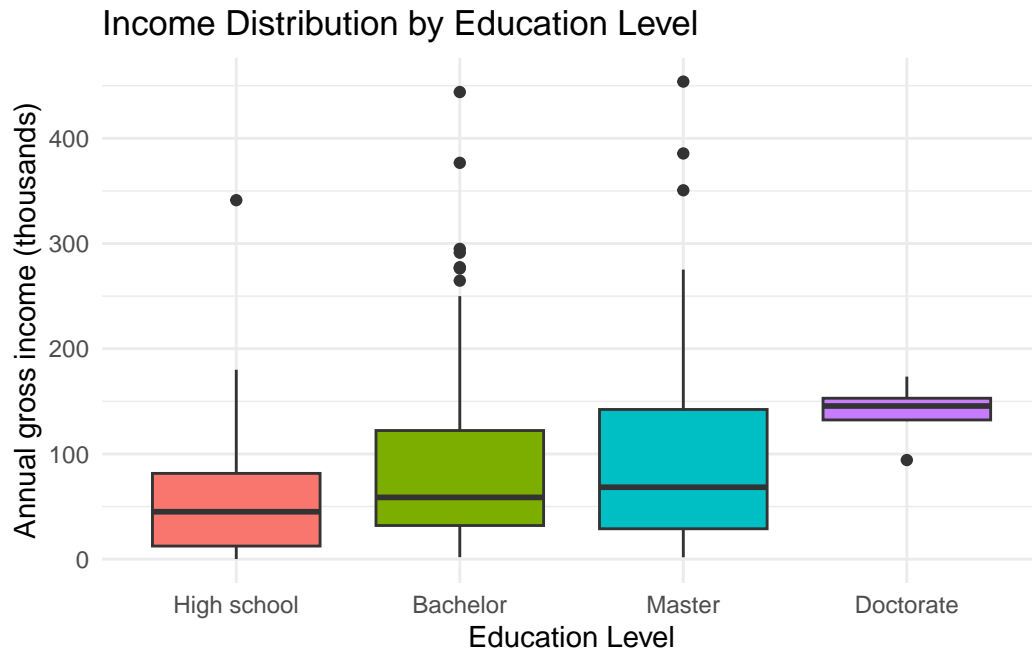
education_group	mean_income	sd_income	n
No University Degree	65.1	72.8	30
University Degree	93.5	89.4	170

University degree holders have a higher mean income and standard deviation, with a difference of \$28,400 in mean compared to non-university degree holders.

Visualizing Income Distribution by Education Level

```
# Order education levels
data$education <- factor(data$education,
                        levels = c("High school",
                                   "Bachelor",
                                   "Master",
                                   "Doctorate"))

# Boxplot comparing income across all education levels
ggplot(data, aes(x = education, y = income, fill = education)) +
  geom_boxplot() +
  theme_minimal() +
  ggtitle('Income Distribution by Education Level') +
  xlab('Education Level') +
  ylab('Annual gross income (thousands)') +
  guides(fill = "none")
```



The boxplot showed that the median of annual gross income increases as the education level gets higher. While there is a huge gap between median income of master and doctorate degree holders, the range of doctorate's income is much narrower, and the maximum point is lower than the other 3 education groups as well.

Confidence Interval for the Difference in Mean Incomes

To compare the difference in income of customer who hold a university degree or not, t-test will be conducted to calculate the confidence interval for the difference in mean incomes.

```
# Subset data into two groups
university_graduates <- data |> filter(education_group == "University Degree")
non_graduates <- data |> filter(education_group == "No University Degree")

# Calculate the mean difference and confidence interval
mean_diff <- mean(university_graduates$income) - mean(non_graduates$income)

# Perform a t-test to get the confidence interval
t_test_result <- t.test(university_graduates$income, non_graduates$income, var.equal = FALSE)

# Print the results
kable(round(mean_diff, 1),
```



```
caption = "Difference in mean income (thousands)",
col.names = "")
```

Table 6: Difference in mean income (thousands)

28.4

```
kable(round(t_test_result$conf.int, 1),
caption = "95% confidence interval of difference in mean income (thousands)",
col.names = "")
```

Table 7: 95% confidence interval of difference in mean income (thousands)

-1.7
58.5

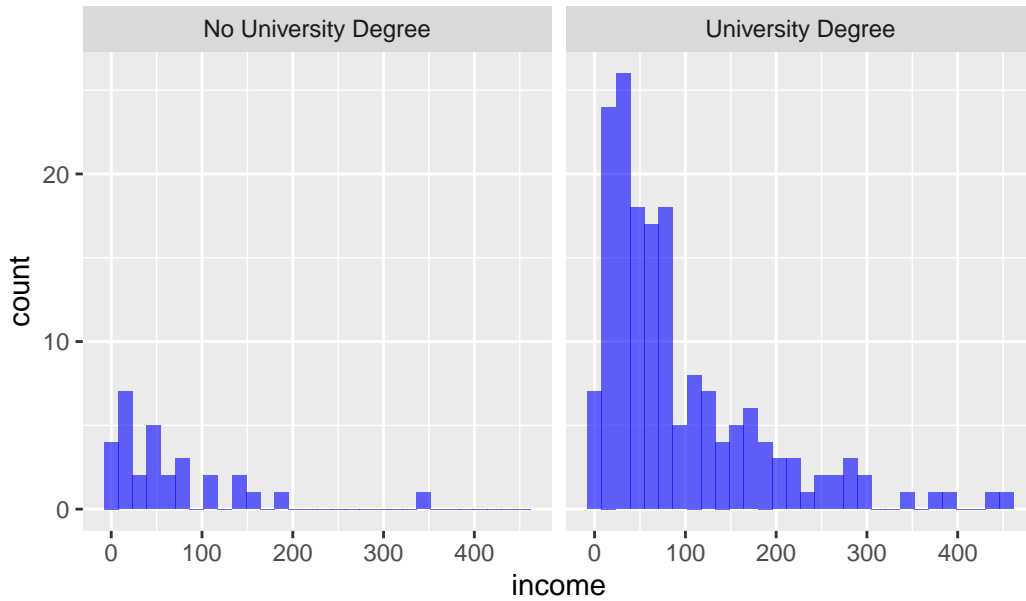
The difference in mean of annual gross income of customers who hold a university degree or not is \$28,400, with a 95% confidence interval of -\$1,700 and \$58,500.

Central Limit Theorem

This part discusses whether it is appropriate to use the Central Limit Theorem when calculating the above confidence interval.

```
# Histograms to check normality
ggplot(data, aes(x = income)) +
  geom_histogram(bins = 30, fill = 'blue', alpha = 0.6) +
  facet_wrap(~education_group) +
  ggtitle('Income Distribution by Education Group')
```

Income Distribution by Education Group



```
# Check the sample sizes of each group
table(data$education_group) %>%
  kable(col.names = c("Education level", "Count"),
        caption = "Sample size of education groups")
```

Table 8: Sample size of education groups

Education level	Count
No University Degree	30
University Degree	170

To use t-test to calculate confidence interval, the sample needs to be normally distributed. But according to the above plots, both groups are not normally distributed. In this case, Central Limit Theorem is needed, which states that the sample mean gets closer to normal distribution as the sample size increases, even if the data are not under normal distribution. Since both groups have a sample size of at least 30, Central Limit Theorem is applicable and t-test could still be used to calculate confidence interval.

Causation of Education and Income

This section of analysis about education level and income seem to show that the average income would be higher as education level increases. However, these are not evidences of a

causal effect. As shown in the t-test results that the confidence interval showed a large positive difference, this could imply that income has a positive relationship with education level, but it could not be summarized with causation that providing greater education would lead to higher incomes.

There could be other underlying factors that are not revealed in this analysis, such as people with higher income have adequate financial support for education, or people in higher level positions are likely to be required by company to complete certain level of education. A positive relationship does not imply causation.