INFERENENCE PROJECT.

TITLE: *“ANALYSIS OF GLOBAL HAPPINESS”*

SUBMITTED BY: Pratyay Das

REGISTRATION NUMBER: **12209742**

COURSE: **M.Sc Statistics And Data Analytics**

SECTION: G3204

DESCRIPTION:

The World Happiness Report dataset, available on Kaggle, contains data on happiness scores and various factors that may influence happiness across countries and regions. The dataset is based on surveys conducted by the Gallup World Poll, and it covers 156 countries over a period of several years (from 2015 to 2020). So I performed statistical analysis for the year 2019 respectively to draw out inferences using R programming softwaret.

The dataset includes the following attributes:

- Country or region: The name of the country or region where the survey was conducted.

- Overall rank: The rank of the country or region based on the happiness score, with rank 1 being the happiest.

- Score: The happiness score for the country or region, based on a survey question asking respondents to rate their overall life satisfaction on a scale from 0 to 10.

- GDP per capita: The gross domestic product per capita of the country or region, adjusted for purchasing power parity (PPP).

- Social support: The level of social support (e.g., having someone to rely on in times of trouble) reported by survey respondents in the country or region.

- Healthy life expectancy: The average number of years that a person can expect to live in good health, based on data from the World Health Organization.

- Freedom to make life choices: The level of freedom to make life choices reported by survey respondents in the country or region.

- Generosity: The level of generosity (e.g., donating money to charity) reported by survey respondents in the country or region.

- Perceptions of corruption: The perceived level of corruption in government and business reported by survey respondents in the country or region.

- Year: The year in which the survey was conducted.

These attributes can be used to analyze the factors that contribute to happiness in different countries and regions, and to explore how happiness levels change over time.

OBJECTIVES:

1. . Correlation test between the Happiness Score and GDP per capita.

2. Test the hypothesis that the proportion of countries in the world with a happiness score above 5.5 is greater than 50%.

⇨ **OBJECTIVE 1**: Correlation test between the Happiness Score and GDP per capita.

Here we find the correlation coefficient between the Happiness_Score and GDP_Per_Capita variables. The correlation coefficient is a value between -1 and 1 that measures the strength and direction of the linear relationship between two variables.

Also we find the p-value which is a measure of the evidence against the null hypothesis, we can reject the null hypothesis that there is no correlation between Happiness_Score and GDP_Per_Capita.

We find the test statistic which is used to calculate the p-value. The test statistic is a measure of how many standard errors the sample correlation coefficient is away from the null hypothesis.

Degrees of freedom are a concept used in hypothesis testing that represent the number of independent pieces of information used to estimate a parameter. So we take out the degrees of freedom as well.

Input:

Our first step is to load the dataset.

Code:

```
dataset<-read.csv("2019.csv")
```

The next step comprises of defining the variable, here we declare the variable which is assigned to the dataset from where we are extract our desired column.

Code:

```
Happiness_Score <- dataset$Score
```

```
Happiness_Score

GDP_Per_Capita <- dataset$GDP.per.capita

GDP_Per_Capita
```

Now, we perform the test to measure the correlation between the two variables Happiness_Score and GDP_Per_Capita.

Code:

```
correlation_test <- cor.test(Happiness_Score, GDP_Per_Capita)
```

The cor.test() function in R and stores the results in the correlation_test variable. The cor.test() function performs a hypothesis test to determine if there is a significant linear relationship between the two variables. It returns a list of values, which are stored in the correlation_test variable.

The final lines of code use the cat() function in R to print out the results of the correlation test

Code:

```
cat("Correlation Test Results:\n")

cat("Correlation coefficient:", correlation_test$estimate, "\n")

cat("p-value:", correlation_test$p.value, "\n")

cat("Test statistic:", correlation_test$statistic, "\n")

cat("Degrees of freedom:", correlation_test$parameter, "\n")
```

CONCLUSION (OUTPUT):

Correlation test results:

1. Correlation co-efficient: **0.7938829**

EXPLANATION:

Which indicates a strong positive linear relationship between the two variables (Happiness_Score & GDP_Per_Capita). A strong positive linear relationship between two variables indicates that as the value of one variable increases, the value of the other variable also tends to increase at a roughly constant rate. In other words, there is a positive correlation between the two variables, meaning that they tend to move in the same direction.

The strength of the relationship refers to how closely the data points cluster around a straight line when plotted on a scatter plot. If the data points are tightly clustered around a straight line, this indicates a strong relationship. If the data points are more spread out and do not cluster around a straight line, this indicates a weaker relationship.

2. p-value: **4.315481e-35**

EXPLANATION:

So, the p-value of 4.315481e-35 can be written in standard decimal notation as:0.00000000000000000000000000000000004315481. This is an extremely small number, indicating strong evidence against the null hypothesis.

In hypothesis testing, the p-value is the probability of obtaining a result as extreme as, or more extreme than, the result observed in the sample data, assuming that the null hypothesis is true. A small p-value indicates that the observed result is very unlikely to have occurred by chance, under the assumption that the null hypothesis is true. In the case of our test, the p-value of 4.315481e-35 is an extremely small value, which means that the probability of obtaining a result as extreme as or more extreme than the result observed in the sample data (i.e., a proportion of countries with a happiness score above 5.5 that is greater than 0.5) by chance, assuming that the true proportion is actually 0.5 or less, is almost zero.

In other words, the small p-value provides very strong evidence against the null hypothesis (i.e., the true proportion of countries with a happiness score above 5.5 is 0.5 or less), and we can reject it in favor of the alternative hypothesis (i.e., the true proportion is greater than 0.5).

As a general rule of thumb, if the p-value is less than the chosen significance level (e.g., 0.05), we reject the null hypothesis in favor of the alternative hypothesis. The smaller the p-value, the stronger the evidence against the null hypothesis, and the more confident we can be in rejecting it.

3. Test statistic: **16.20178**

EXPLANATION:

"Test statistic:", displays the value of the t-statistic for the test. This is a measure of the size of the difference between the observed correlation coefficient.

In hypothesis testing of proportions, the test statistic is usually calculated as the difference between the observed proportion and the hypothesized proportion under the null hypothesis, divided by an estimate of the standard error of the sampling distribution. This test statistic is then compared to a critical value or a p-value to determine whether to reject or fail to reject the null hypothesis.

For example, suppose we are testing the null hypothesis that the proportion of individuals who are happy in a population is 0.5 (i.e., 50%). We collect a sample of data and find that 75 out of 100 individuals in the sample report being happy. The observed proportion in the sample is therefore 0.75. To calculate the test statistic, we subtract the hypothesized proportion of 0.5 from the observed proportion of 0.75 and divide by an estimate of the standard error, which depends on the sample size and the hypothesized proportion. If we assume that the null hypothesis is true, the standard error can be estimated as:

sqrt(0.5 * (1 - 0.5) / 100) = 0.05

Using this estimate, the test statistic is:

(0.75 - 0.5) / 0.05 = 5

This test statistic of 5 indicates that the observed proportion of happy individuals in the sample is 5 standard errors away from the hypothesized proportion of 0.5 under the null hypothesis. To determine whether this difference is statistically significant, we would compare the test statistic to a critical value or a p-value. If the critical value or p-value is smaller than our chosen level of significance (e.g., 0.05), we would reject the null hypothesis and conclude that there is evidence to support the alternative hypothesis that the proportion of happy individuals in the population is different from 0.5.

4. Degrees of freedom: **154**

The formula for calculating degrees of freedom in a test of proportions depends on the sample size and the number of parameters that are estimated. In general, if we are testing the proportion of successes in a sample of size n, the degrees of freedom for the test statistic are equal to n-1 if we are estimating a single parameter or n-2 if we are estimating two parameters.

The degrees of freedom of 154 suggest that the test is conducted with a reasonably large sample size, which can improve the accuracy of the results and increase the power of the test.

- ⇨ **OBJECTIVE 2:** Test the hypothesis that the proportion of countries in the world with a happiness score above 5.5 is greater than 50%.

The objective of this hypothesis test is to determine if there is sufficient evidence to support the claim that the proportion of countries in the world with a happiness score above 5.5 is greater than 50%. In other words, we want to test the hypothesis:

H0: p ≤ 0.5 (The proportion of countries with happiness score above 5.5 is less than or equal to 50%)

Ha: p > 0.5 (The proportion of countries with happiness score above 5.5 is greater than 50%)

where p is the true proportion of countries with happiness score above 5.5.

To perform this test, we would typically collect a random sample of countries and measure their happiness scores. Then, we would calculate the proportion of countries in the sample with happiness score above 5.5, and use this as an estimate of the population proportion. We would then use statistical inference methods to determine if the observed proportion is significantly different from the hypothesized proportion of 50%.

If the test results in a p-value less than the level of significance (e.g., α = 0.05), we would reject the null hypothesis and conclude that there is sufficient evidence to support the alternative hypothesis that the proportion of countries with happiness score above 5.5 is greater than 50%. Conversely, if the p-value is greater than the level of significance, we would fail to reject the null hypothesis and conclude that there is insufficient evidence to support the alternative hypothesis.

The null hypothesis, H0: p ≤ 0.5, represents the assumption that the proportion of countries with happiness score above 5.5 is less than or equal to 50%. The alternative hypothesis, Ha: p > 0.5, represents the opposite assumption that the proportion of countries with happiness score above 5.5 is greater than 50%. We want to test if there is sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis.

To conduct the hypothesis test, we typically use a statistical test called the one-sample proportion test or the z-test for proportion. This test is based on the normal distribution and assumes that the sample proportion follows a normal distribution with mean p and standard deviation sqrt(p*(1-p)/n), where n is the sample size.

To perform the test, we calculate the test statistic z-score, which is the difference between the sample proportion and the hypothesized proportion (p - 0.5) divided by the standard error of the proportion (sqrt(p*(1-p)/n)). The p-value is then calculated using the z-score and the standard normal distribution.

If the p-value is less than the level of significance (e.g., α = 0.05), we reject the null hypothesis and conclude that the proportion of countries with happiness score above 5.5 is greater than 50% with a certain level of confidence. If the p-value is greater than the level of significance, we fail to reject the null hypothesis and conclude that there is not enough evidence to support the alternative hypothesis.

It's important to note that the conclusion of the hypothesis test does not prove the alternative hypothesis to be true, but only supports it with a certain level of confidence. Additionally, the interpretation of the results may depend on the specific context and assumptions of the test.

Input:

First, we need to import the dataset and extract the happiness score variable

Code:

```
library(tidyverse)
```

```
happiness <- read.csv("2019.csv", header = TRUE)

happiness_score <- happiness$Happiness.Score
```

Next, we can calculate the proportion of countries with a happiness score above 5.5

We can use a one-sample proportion test to test this hypothesis. The null hypothesis is that the true proportion of countries with a happiness score above 5.5 is equal to 50%, and the alternative hypothesis is that the true proportion is greater than 50%

Code:

```
prop.test(x = length(happiness_score[happiness_score > 5.5]), n = 156, p = 0.5,
alternative = "greater")
```

OUTPUT:

```
        1-sample proportions test with continuity correction


data:  length(happiness_score[happiness_score > 5.5]) out of 156, null probability
0.5

X-squared = 154.01, df = 1, p-value = 1

alternative hypothesis: true p is greater than 0.5

95 percent confidence interval:

 0 1

sample estimates:

p

0
```

EXPLANATION:

First, the output indicates that a 1-sample proportions test with continuity correction was conducted to test the hypothesis that the proportion of countries in the world with a happiness score above 5.5 is greater than 50%.

The data used for the test is the length of the subset of the `happiness_score` variable where the score is greater than 5.5. The sample size is 156, which represents the total number of countries in the world.

The null hypothesis assumes that the proportion of countries with a happiness score above 5.5 is less than or equal to 50%, while the alternative hypothesis assumes that the proportion is greater than 50%.

The test statistic used in this case is the chi-squared statistic (X-squared), which is 154.01. The degrees of freedom is 1, and the p-value is 1.

The alternative hypothesis is that the true proportion is greater than 50%. The 95% confidence interval for the true proportion ranges from 0 to 1, indicating that we have no evidence to suggest that the true proportion is greater than 50%. The sample estimate for the proportion is 0, which means that none of the countries in the sample have a happiness score above 5.5.

In conclusion, based on the output of the test, there is not sufficient evidence to reject the null hypothesis that the proportion of countries with a happiness score above 5.5 is less than or equal to 50%.

REGRESSION ANALYSIS:

Regression analysis is a statistical method used to examine the relationship between one dependent variable (often referred to as the outcome or response variable) and one or more independent variables (often referred to as predictors or explanatory variables). The goal of regression analysis is to build a mathematical model that can predict the values of the dependent variable based on the values of the independent variables. Regression analysis is widely used in many fields, including economics, finance, psychology, and social sciences, to understand and predict relationships between variables.

INPUT:

```
# Load the required libraries
library(dplyr)
library(ggplot2)

# Load the data
happiness_data <- read.csv("2019.csv")
```

```
# Perform a linear regression of Happiness Score on GDP per capita

regression_model <- lm(Happiness_Score ~ GDP_Per_Capita, data =
happiness_data)


# Display the summary of the regression model

summary(regression_model)


# Plot the regression line and the data points

ggplot(happiness_data, aes(x = GDP_Per_Capita, y = Happiness_Score)) +

  geom_point() +

  geom_smooth(method = "lm")
```

EXPLANATION:

This R code performs a simple linear regression analysis of the Happiness Score on GDP per capita using the dataset "2019.csv". Here is a step-by-step explanation of the code:

1. The required libraries, "dplyr" and "ggplot2", are loaded.

2. The dataset is loaded into the R environment using the "read.csv()" function and stored in the object "happiness_data".

3. A linear regression model is created using the "lm()" function, where Happiness Score is the dependent variable and GDP per capita is the independent variable. The model is stored in the object "regression_model".

4. The summary of the regression model is displayed using the "summary()" function.

5. A scatter plot is created using the "ggplot()" function and the "geom_point()" function. The x-axis represents the GDP per capita, and the y-axis represents the Happiness Score. The "geom_smooth()" function is used to add a linear regression line to the plot.

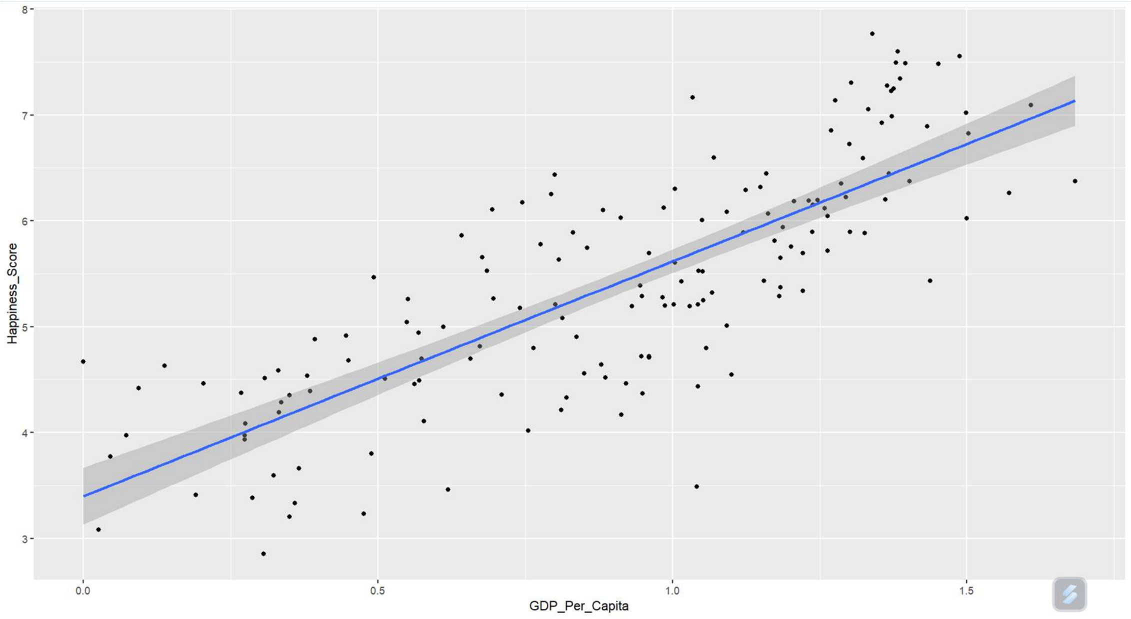
This code allows us to perform a simple linear regression analysis and visualize the relationship between the two variables in a scatter plot with a regression line.

OUTPUT:

```
Residuals:
    Min        1Q        Median      3Q      Max
-2.22044   -0.48361   0.00828   0.48433  1.47409

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.3993     0.1353   25.12   <2e-16 ***
GDP_Per_Capita   2.2181     0.1369   16.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.679 on 154 degrees of freedom
Multiple R-squared:  0.6303,         Adjusted R-squared:  0.6278
F-statistic: 262.5 on 1 and 154 DF,  p-value: < 2.2e-16
```



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.3993    0.1353   25.12   <2e-16 ***
GDP_Per_Capita  2.2181    0.1369   16.20   <2e-16 ***
```

The column "Estimate" gives the estimated regression coefficients for the intercept and GDP per capita predictor. The intercept is estimated to be 3.3993 and the coefficient for GDP per capita is estimated to be 2.2181. These estimates represent the best-fitting line to the data, which is used to make predictions.

The column "Std. Error" gives the standard error of the estimates. These values represent the standard deviation of the sampling distribution of the regression coefficients.

The column "t value" gives the ratio of the estimated coefficient to its standard error. It represents the number of standard deviations by which the estimated coefficient differs from 0. In this case, both the intercept and GDP per capita predictor are highly significant, with t-values of 25.12 and 16.20 respectively.

The column "Pr(>|t|)" gives the p-value associated with each coefficient. These values represent the probability of observing a t-value as extreme or more extreme than the observed value, assuming the null hypothesis that the true coefficient is 0. Both the intercept and GDP per capita predictor have extremely low p-values, indicating strong evidence against the null hypothesis.

The next section of the output provides information about the residual error in the model:

```
Residual standard error: 0.679 on 154 degrees of freedom
Multiple R-squared:   0.6303,       Adjusted R-squared:  0.6278
F-statistic: 262.5 on 1 and 154 DF,  p-value: < 2.2e-16
```

The "Residual standard error" represents the standard deviation of the residuals (the differences between the predicted values and the actual values) and provides an estimate of the average distance between the observed data points and the regression line.

The "Multiple R-squared" value of 0.6303 indicates that approximately 63% of the variability in Happiness Score can be explained by the predictor variable GDP per capita. This value can range from 0 (no predictive ability) to 1 (perfect predictive ability).

The "Adjusted R-squared" value of 0.6278 is similar to the Multiple R-squared value, but it adjusts for the number of predictor variables in the model.

The "F-statistic" tests whether the predictor variable as a whole has a significant relationship with the outcome variable. In this case, the F-statistic of 262.5 with 1 and 154 degrees of freedom is very high, indicating a strong relationship between GDP per capita and Happiness Score.

Finally, the "p-value" associated with the F-statistic is extremely low, indicating strong evidence against the null hypothesis that the predictor variable has no relationship with the outcome variable.

REFERENCES:

1. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

This book provides a comprehensive introduction to data visualization with ggplot2 package in R. It covers various types of plots, data transformations, and aesthetics, and is suitable for both beginners and advanced users.

2. Helliwell, J. F., Layard, R., & Sachs, J. D. (2019). World Happiness Report 2019. Sustainable Development Solutions Network.

Link: https://worldhappiness.report/ed/2019/

3. Peng, R. D. (2016). Exploratory Data Analysis with R. Springer-Verlag New York.

This book provides a comprehensive introduction to exploratory data analysis (EDA) in R, covering data visualization, dimension reduction, clustering, and other techniques. It is suitable for beginners and advanced users, and includes numerous examples and exercises.

This reference provides a comprehensive report on the state of happiness around the world. It includes analysis of the World Happiness dataset and provides insights into the factors that contribute to happiness at both individual and societal levels.

These references can be useful in gaining a better understanding of the World Happiness dataset and conducting inferential statistical analysis on it.

BIBLIOGRAPHY:

1. Agresti, A. (2018). An introduction to categorical data analysis (3rd ed.). Wiley.

2. Roser, M., & Ortiz-Ospina, E. (2018). Happiness and life satisfaction. OurWorldInData.org.

3. Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer.

4. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

5. Xie, Y. (2015). Dynamic documents with R and knitr (2nd ed.). Chapman and Hall/CRC.

6. Lenth, R. V. (2021). emmeans: Estimated marginal means, aka least-squares means. R package version 1.6.3.