ISSN: 1813-4890

Analysis of Travel Behavior Based on Gender

Chenjia Liu

School of North China University of Technology, Beijing, 100144, China

liuchenjia1002@163.com

Abstract

Tourism companies are now constantly seeking to provide services that meet the individualized and customized needs of tourists. Gender is an important consideration that influences many of these issues. This report explores whether gender has a significant impact on tourism through a comparative analysis of males and females in terms of tourism visits and spend, in the hope of providing some insights into the development of the tourism industry. In this report, firstly, all the data were pre-processed, and then all the data were screened according to the research objectives and the principle of controlling a single variable, and finally 406 data were analyzed by grouping the variables VISITS and Spend according to gender. Firstly, the overall characteristics of the data were observed through descriptive analysis, then the population means of different variables for different genders were estimated by using the central limit theorem. Finally, based on the results of the previous analysis, a reasonable hypothesis was made. After the normality test, the original and transformed data did not pass the test, so the Mann-Whitney U test was selected for the significance test, it was found that Visits and Spend did have significant differences in terms of gender.

Keywords

Gender Differences, Tourism Spend, Tourism Visits, Mann-Whitney U Test.

1. Introduction

1.1. Background

Tourism is one of the world's most important economic sectors, employing one in ten people on the planet and providing livelihoods for hundreds of millions of people. In some countries, tourism revenues can account for more than 20 per cent of their gross domestic product (GDP). Therefore, the study of the tourism industry is very important. At the same time, tourism, as a collection of individual human practices, behaviors and activities, its construction, presentation and consumption are always gendered. By studying the influence of gender on tourism behavior, it helps the tourism industry to develop more precisely and to create personalized tourism services according to gender.

1.2. Data Selection

The datasets are produced by International Passenger Survey Overseas Travel and Tourism Data Sets guide. The categorical variable selected for this report is SEX and the continuous variables are VISITS and SPEND. In order to exclude the influence of other variables as much as possible, this report only analyzes the Travelpac dataset with a sample size of 1 in 2023, removes missing values, and ultimately integrates 406 data, which are divided into two groups according to gender, with 242 males and 164 females.

1.3. Research Questions

Objective 1: Understanding data characteristics through descriptive analysis

ISSN: 1813-4890

Objective 2: Using the Central Limit Theorem to estimate confidence intervals for the population mean.

Objective 3: Using hypothesis test to analyze whether gender has a significant effect on VISITS and SPEND.

2. Literature Review

Dae-Young Kim et al. in the paper “Gender differences in online travel information search: Implications for marketing communications on the internet”. They found that Gender has been and continues to be one of the most common forms of segmentation used by marketers in general and advertisers in particular. In general, males and females are likely to differ in information processes and decision making.[1]

Wafa Elias et al. in the paper “Gender and Travel Behavior in Two Arab Communities in Israel”. They used Hypothesis Test to analysis the critical but understudied issue of gender differences in travel behaviors in traditional societies, in general, and in the Arab world, in particular. [2]

Freund Daniela et al in the paper "Exploring the gendered tourism entrepreneurial ecosystem in Barcelona and responses required: a feminist ethic of care". They foucued on gender biases and opportunities to enhance equity in the travel ecosystem in Barcelona by employing an ethic of care and committing to diversification.[3]

The above literature illustrates that gender does have an impact on some preferences in travel and that measures have begun to be taken to optimize the user experience for gender differences and to promote further development of the tourism industry. It also shows that the research on gender differences on the number of visits and spending is necessary and reasonable.

3. Methodology

3.1. Descriptive Analysis

Descriptive statistical analysis is a method of data analysis aimed at summarizing and describing data to better understand the characteristics and trends of the data. It presents the basic characteristics of the data, including central tendency, degree of dispersion, distribution pattern, etc., mainly through statistical indicators and charts.

Firstly, a descriptive analysis of the data under study was carried out by using MATLAB to calculate the mean, median, mode, standard deviation, sample variance, range, minimum and maximum for each set of data. Then the frequency distribution histogram of each group of data was plotted to observe the distribution characteristics.

3.2. Confidence Interval

A confidence interval is a range of values that is likely to contain a population parameter with a certain level of confidence. In this report, the population standard deviation is unknown. However, the sample size is large enough ($n = 40 \geq 30$). Therefore, a standard normal distribution (Z-distribution) was used to calculate the confidence intervals. The formula is as follows.

$$\text{Confidence Interval} = \bar{X} \pm Z_{\frac{\alpha}{2}} * \frac{S}{\sqrt{n}}$$

3.3. Central Limit Theory

The central limit theorem states that the sampling distribution of a sample mean is approximately normal if the sample size is large enough ($n \geq 30$), even if the population

ISSN: 1813-4890

distribution is not normal. In this report, I randomly selected 40 data at a time from the selected samples as a sample to calculate each sample mean, for a total of 100 randomly selected samples, and used the sample means to estimate the confidence interval in which the population mean lies. It is hoped that the population mean confidence intervals will be used to make reasonable hypotheses.

3.4. Hypothesis Test

Hypothesis testing is a tool for making statistical inferences about population data. It is an analysis tool that tests assumptions and determines how likely something is within a given standard of accuracy. Hypothesis testing provides a way to verify whether the results of an experiment are valid.

Based on the characteristics of the data obtained from the descriptive analysis and the population confidence interval, we formulate the null and alternative hypotheses and perform Mann-Whitney U test on the data of both men and women to verify if there is a significant difference.

4. Data Analysis

4.1. Descriptive Analysis by Gender Grouping

Table 1 Descriptive Analysis of Visits

Male Visits		Female Visits	
Mean	2541.92869	Mean	2516.327806
Median	1873.012	Median	2315.48527
Mode	1092.939	Mode	1092.93936
Standard Deviation	1896.85882	Standard Deviation	1829.43746
Sample Variance	3598073.383	Sample Variance	3346841.42
Range	12976.713	Range	15680.32616
Minimum	531.921	Minimum	569.934845
Maximum	13508.634	Maximum	16250.261
Count	242	Count	164

Females exhibit higher median and lower Mean for the number of visits, but the data have less variability and a more concentrated distribution; Males have a wider distribution of visit counts, greater variability, and longer tails in the data distribution, with the possibility of some individuals with extreme high visit counts.

Table 2 Descriptive Analysis of Spend

Male Spend		Female Spend	
Mean	1132098.664	Mean	1073982.079
Median	754521.362	Median	679450.1374
Mode	109293.936	Mode	278699.5368
Standard Deviation	1134731.54	Standard Deviation	1216933.371
Sample Variance	1.28762E+12	Sample Variance	1.48093E+12
Range	6841931.943	Range	7185401.167
Minimum	10782.275	Minimum	32788.1808
Maximum	6852714.218	Maximum	7218189.348
Count	242	Count	164

Males exhibit higher mean and median spend, but the data have less variability and a more concentrated distribution; Females, although they have lower mean spend, have greater

ISSN: 1813-4890

variability and a wider range of spend and more pronounced spikes in the distribution of the data.

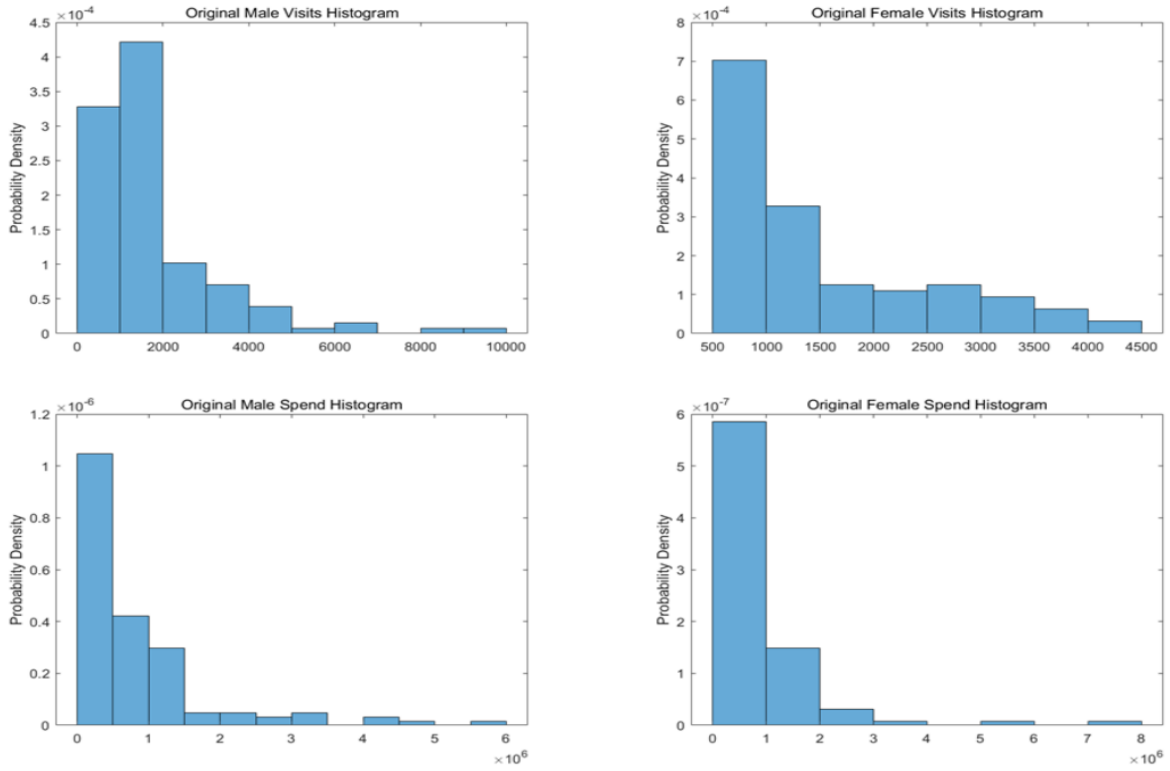

Fig.1 Histogram of the Frequency Distribution of the Four Data Sets

Through the frequency distribution histogram above we can see that all four sets of data show a right-skewed distribution. However, the data analysis process often requires the data to be normally distributed, which means that we need to perform some transformations on the data at a later stage, such as log-transformation, or use analytical methods that do not require normal distribution.

4.2. Sample Means Estimate Population Mean

For each set of data, we randomly sample 100 times with a size of 40 and calculate the mean of each sample, resulting in 100 sample means. The frequency distribution histogram and Q-Q plot show that the sample means are perfectly normally distributed.

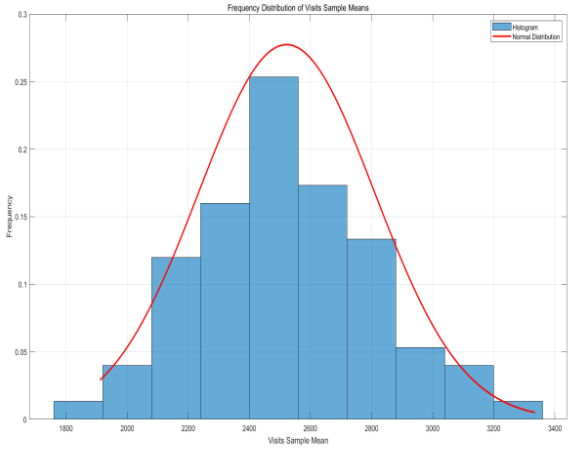

Fig.2 Frequency Distribution of Visits Sample Mean

ISSN: 1813-4890

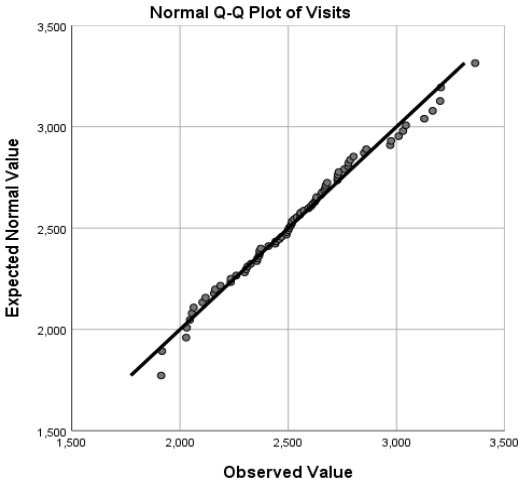

Fig.3 Normal Q-Q Plot of Visits Sample Mean

Next, the population mean under the corresponding 95% confidence interval for each set of data was then calculated using the following formula.

$$\text{Confidence Interval} = \bar{X} \pm Z_{\frac{\alpha}{2}} * \frac{S}{\sqrt{n}}$$

For 95% confidence level, $Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$. Thus, the population mean confidence intervals corresponding to the four data sets are as follows.

Table 3 Confidence Interval for Population Mean of Four groups

	95% Confidence Interval for Population Mean
Male Visits	[1534.69,1587.94]
Female Visits	[1519.34,1589.00]
Male Spend	[1527664.38,1647097.53]
Female Spend	[1390698.52,1497837.25]

These confidence intervals provide us with statistical estimates of the number of tourism visits and spend by gender. The overlap of the confidence intervals does not mean that there are no differences between the two groups, in order to further determine if there are statistically significant differences, hypothesis tests such as two sample t-tests or Mann-Whitney U-tests can be conducted.

4.3. Gender Differences Have an Impact on Travel Behavior

Generally, a two-sample t-test is used to test whether there is a significant difference between two independent sample distributions, however, neither the original nor the transformed data passed the normality test. (Only Log-Transformed Female Spend passed the normality test.) Therefore, the Mann-Whitney U test is taken to conduct the test.

Table 4 Test of Normality

	Kolmogorov-Smirnova			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Male Visits Raw Data	0.172	242	3.9816E-19	0.766	242	3.2467E-18
Male Visits Log-Transformed Data	0.116	242	2.3195E-8	0.965	242	0.000014
Male Visits Sqrt-Transformed Data	0.146	242	1.9931E-13	0.894	242	5.9559E-12
Male Spend Raw Data	0.184	242	4.3741E-22	0.794	242	4.0781E-17
Male Spend Log-Transformed Data	0.082	242	0.000533	0.962	242	0.000006
Male Spend Sqrt-Transformed Data	0.096	242	0.000012	0.953	242	5.4343E-7

ISSN: 1813-4890

Female Visits Raw Data	0.185	164	5.9487E-15	0.709	164	1.2339E-16
Female Visits Log-Transformed Data	0.087	164	0.004151	0.963	164	0.000239
Female Visits Sqrt-Transformed Data	0.111	164	0.000036	0.887	164	7.4805E-10
Female Spend Raw Data	0.206	164	6.4558E-19	0.696	164	5.4795E-17
Female Spend Log-Transformed Data	0.051	164	0.2	0.993	164	0.649
Female Spend Sqrt-Transformed Data	0.122	164	0.000003	0.905	164	8.0704E-9

The Mann-Whitney U test is a non-parametric test used to determine whether there is a significant difference between the distributions of two independent samples. It is often used as an alternative to the t-test when the data does not satisfy the assumption of normality. The Mann-Whitney U test is achieved mainly through the following steps:

Step 1: State the null and alternate hypothesis

The null hypothesis states that the median difference between the pairs ranks of the observations is zero (that is there is no difference in the ranks of the two pairs of observations) and the alternate hypothesis states that the median difference between the ranks of data is not zero.

$$H_0: \mu_M = \mu_F$$
$$H_1: \mu_M \neq \mu_F$$

Step 2: Compute the rank sum

First, the two samples are combined and sorted, and then a rank is assigned to each element. If there are equal values, assign an average rank. For each sample, calculate the rank sum of all its elements.

Step 3: Compute the U statistics

The U statistic is the centerpiece of the Mann-Whitney U test. For sample X, the U statistic is calculated as:

$$U_x = R_x - \frac{n_x(n_x + 1)}{2}$$

Where R_x is the rank sum of sample X, n_x is the sample size of X. Similarly for sample Y:

$$U_y = R_y - \frac{n_y(n_y + 1)}{2}$$

The smaller of U_x and U_y is then taken as the final U statistic.

Step 4: Compute the expected value of U and standard Deviation of U

Under the null hypothesis (two samples from the same distribution), the expected value of the U statistic is:

$$E(U) = \frac{n_x n_y}{2}$$

The standard deviation of the U statistic is:

$$\sigma_U = \sqrt{\frac{n_x n_y (n_x + n_y + 1)}{12}}$$

Step 5: Compute Z statistics

When the sample size is large (usually considered to be greater than 30 per group), the U statistic can be approximated to follow a normal distribution. The z-value is calculated using the following formula:

ISSN: 1813-4890

$$Z = \frac{U - E(U)}{\sigma_U}$$

Step 6: Compute P values

Calculate the p-value using the following equation:

$$p = 2 * (1 - \Phi(|Z|))$$

Where Φ is the CDF(Cumulative Distribution Function) of the standard normal distribution.

Following the above steps, the Mann-Whitney U test results for Visits and Spend on gender grouping were calculated using MATLAB as follows.

Table 5 Ranks of Visits and Spend

	Sex	N	Mean Rank	Sum of Ranks
Spend	Male	242	169.99	41137
	Female	164	252.95	41484
	Total	406		
Visits	Male	242	166.62	40322
	Female	164	257.92	42299
	Total	406		

Table 6 Test Statistics for Visits and Spend

Test Statistics		
	Visits	Spend
Mann-Whitney U	10919	11734
Z	-7.693	-6.99
P-Value (2-tailed)	1.4337E-14	2.7464E-12

The significance level α was set to 0.05. The p-value for both sets of data is much less than 0.05, so based on the two-tailed test it can be concluded that the H_0 hypothesis can be rejected at the 95% confidence interval that there is a significant difference between males and females in terms of travel VISITS and SPEND.

5. Conclusion

5.1. Strengths

In this paper, the data were first pre-processed scientifically, including the elimination of outliers and reasonable random sampling according to the research objectives. Then various statistical methods such as descriptive analysis, confidence intervals, and hypothesis testing were applied from multiple perspectives to analyze the effect of gender on two important continuous variables in tourism - visits and spending. According to the characteristics of non-normal distribution of data, the appropriate Mann-Whitney U test was chosen for significance analysis. In general, the research ideas in this paper are clear and logical and rigorous reasonable use of statistical methods to analyze and deal with practical problems.

5.2. Weaknesses

The generalizability of the sample needs to be further improved as only data from a certain period was selected, Besides, the test was unable to understand what specific differences and the reasons for the differences in travel preferences by gender there are, these issues need to be further researched.

ISSN: 1813-4890

6. Outlook

This paper demonstrates the impact of gender on tourism visits and consumption. Due to the limited sample selection, more data will be collected in the future to further validate the reliability of the results through different time and different geographical areas. In addition to this, the reasons for the impact will be further analyzed and how to provide better services to tourists based on the differences to promote tourism more effectively. In view of the diverse and variable needs of tourists, timely measures to take the demand of tourists as the guide for the development of tourism products and the layout of service function scenarios, and accurately locating the demand for tourism consumption and services are the prerequisites.

References

[1] D. Kim, Y.X. Lehto, M.A. Morrison: Gender Differences in Online Travel Information Search: Implications for Marketing Communications on the Internet, Tourism Management, Vol. 28 (2007) No.2, p.423-433.

[2] W. Elias, L.G. Newmark, Y. Shiftan: Gender and Travel Behavior in Two Arab Communities in Israel, Transportation Research Record, Vol. 2067 (2008) No.1, p.75-83.

[3] F. Daniela, G.R. Itziar, B.A. K, et al.: Exploring the Gendered Tourism Entrepreneurial Ecosystem in Barcelona and Responses Required: A Feminist Ethic of Care, Journal of Sustainable Tourism, Vol. 32 (2024) No.3, p.637-655.