Data Analysis and Presentation using Statistical Techniques

Neville Greening

Doctor of Philosophy Academic Mixed Research Methods
Osaka University-8-1-1 Aomatani-Higashi, Minoh,
OSAKA 562-8558 JAPAN
Email: Osaka.university2017@gmail.com

DOI: 10.31364/SCIRJ/v7.i9.2019.P0919692
http://dx.doi.org/10.31364/SCIRJ/v7.i9.2019.P0919692

Abstract: This assignment deals with two primary case studies – one dealing with the estimation of several golfers in the U.S.A, and those who rank top in earnings, and the second case study will focus on the hypothesis test for the mean and standard deviation of the temperature of hot chocolates. The data for golfers were collected through observational research, whereas the research data for temperature is collected through experimental research. In this research, both descriptive and inferential statistics are used for the analyses. In both case studies, we have used Minitab as an analytical tool.

Keywords: hypothesis test, standard deviation, descriptive and inferential statistics, Minitab

Introduction

Case 1: This mini-case assignment aims to estimate the number of golfers in American from the age of eighteen and above and to estimate the amount of golfers who earn over $75,000.

Case 2: This assignment aims to verify whether the temperature of hot chocolate at D's Bagels owned by David is maintained at the desired temperature within the desired standard deviation

The two mini-case assignments are successfully analyzed using statistical techniques.

Mini-case assignment-1

This mini-case assignment aims to estimate the number of golfers in American from the age of eighteen and above and to estimate the number of golfers who earn over $75,000. The data was collected through a survey of 1,116 U.S adults of age 18 and above. The responses were recorded as the categories of household income and whether the subject had played golf at least once during summer. The variable household income is structured as an ordinal variable, and the variable golfer is structured as a dichotomous variable. The appropriate graphical methods and descriptive methods to display the above use two variables, as mentioned below. Descriptive analysis refers to the transformation of raw data into a form that will make it easy to understand and interpret (ZIKMUND, William

© 2019, Scientific Research Journal

G., 2003). The most appropriate descriptive method to display categorical data is the cross-tabulation. The following cross-tabulations present the data of golfers categorized by income.

Tabulated statistics: Income, Golfer?

Rows: Income	Columns: Golfer?		
	1	2	All
1	208	6	214
	97.20	2.80	100.00
	22.15	3.39	19.18
	18.64	0.54	19.18
2	169	7	176
	96.02	3.98	100.00
	18.00	3.95	15.77
	15.14	0.63	15.77
3	148	12	160
	92.50	7.50	100.00
	15.76	6.78	14.34
	13.26	1.08	14.34
4	172	30	202
	85.15	14.85	100.00
	18.32	16.95	18.10
	15.41	2.69	18.10
5	189	54	243
	77.78	22.22	100.00
	20.13	30.51	21.77
	16.94	4.84	21.77
6	53	68	121
	43.80	56.20	100.00
	5.64	38.42	10.84
	4.75	6.09	10.84
All	939	177	1116
	84.14	15.86	100.00
	100.00	100.00	100.00
	84.14	15.86	100.00

Cell Contents: Count
% of Row
% of Column

Figure 1: Cross-tabulation – Income and Golfer

Tabulated statistics: Golfer and above $75,000

Rows: Golfer and above $75,000

	Count	% of Row	% of Column	% of Total
0	1048	100	93.91	93.91
1	68	100	6.09	6.09
All	1116	100	100.00	100.00

Figure 2: Tabulation – Golfer earning over $75,000

The appropriate graph to display the categorical data is the bar graph. The following bar graphs show the distribution of golfers and non-golfers across the income categories.

© 2019, Scientific Research Journal
http://dx.doi.org/10.31364/SCIRJ/v7.i8.2019.P0819XX

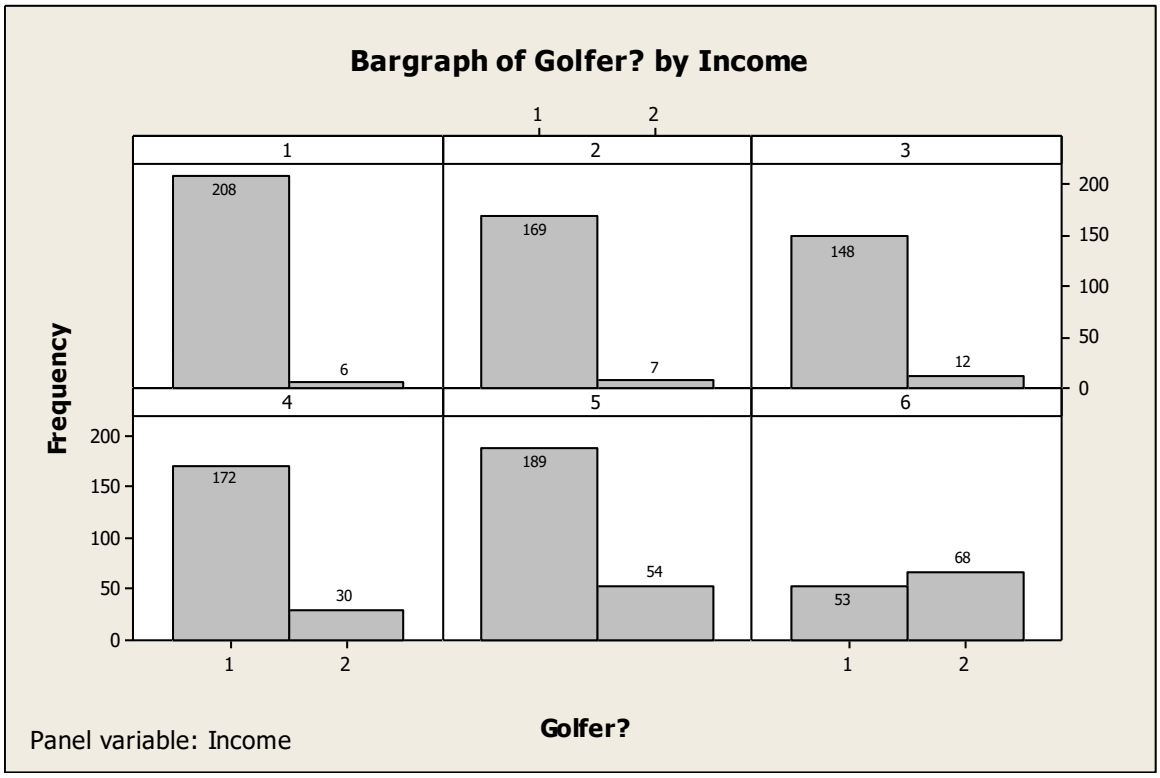

Figure 3: Bar graphs number of non-golfers and golfers by income

In the above graph, the difference in the number of golfers and non-golfers is negligible in the 6th income category, which is over $75,000. The most appropriate measure of central tendency for the income data is mode as it is the only valid measure of central tendency for categorical data.

The following is the Minitab output of descriptive statistics of income categorized by golfers.

Descriptive Statistics: Income

Variable	Golfer?	Mode	N for Mode
Income	1	1	208
	2	6	68

Figure 4: Descriptive statistic of income by Golfers

The above table shows that most non-golfers are of income level under $35,000, and the majority of golfers are of income level over $75,000. This measure somewhat leads to a rough approximation that golf is the rich men's game.

© 2019, Scientific Research Journal
http://dx.doi.org/10.31364/SCIRJ/v7.i8.2019.P0819XX

From the cross-tabulation, it is evident that 177 of the 1,116 persons are golfers. Estimating the total number of golfers is equivalent to calculating the population proportion of golfers from the sample proportion. Evaluating with 95% confidence is framing the 95% confidence interval. The 95% confidence interval for population proportion *p* is represented as follows using the standard normal distribution as the sampling distribution.

$$p \pm z_{0.025}\sqrt{\frac{p\left(1-p\right)}{n}}$$ (VOELKER, David H., and Orton, Peter Z., 2001)

Here p is the sample proportion, $z_{0.025}$ is the two-sided critical value of standard normal distribution at 95% confidence level, and n is the sample size. Their following conditions need to be satisfied for estimating proportions.

$$\begin{aligned} n\,p &\geq 10 \\ n\left(1-p\right) &\geq 10 \end{aligned}$$ (VOELKER, David H., and Orton, Peter Z., 2001)

Substitute n = 1,116 and $p = 0.16$ in the above conditions.

$$\begin{aligned} 1{,}116\left(0.16\right) &\approx 179 > 10 \\ 1{,}116\left(1-0.16\right) &\approx 937 > 10 \end{aligned}$$

The conditions are satisfied. The following Minitab output provides a 95% confidence interval for the proportion of golfers.

CI for One Proportion: Golfer?

Event = 2

Variable	X	N	Sample p	95% CI
Golfer?	177	1116	0.158602	(0.137644, 0.181378)

Figure 5: 95% Confidence interval for the percentage of golfers

From the above Minitab output, the 95% confidence interval for the population proportion of golfers in the U.S is (0.14, 0.18). With 95% confidence, it can be estimated that the percentage of golfers in the U.S is between 14% and 18%.

From the cross-tabulation, it is evident that 68 of the 1,116 persons are golfers earning over $75,000. Estimating the total number of golfers making over $75,000 is equivalent to determining the population proportion of golfers earning over $75,000 from the sample proportion. Calculating with 95% confidence is comparable to framing the 95% confidence interval. The 95% confidence interval for population proportion *p* is represented as follows using the standard normal distribution as the sampling distribution.

© 2019, Scientific Research Journal
http://dx.doi.org/10.31364/SCIRJ/v7.i8.2019.P0819XX

$$p \pm z_{0.025}\sqrt{\frac{p(1-p)}{n}}$$ (VOELKER, David H., and Orton, Peter Z., 2001)

Here p is the sample proportion, $z_{0.025}$ is the two-sided critical value of standard normal distribution at 95% confidence level, and n is the sample size. Their following conditions need to be satisfied for estimating proportions.

$$\begin{aligned} n\,p &\geq 10 \\ n(1-p) &\geq 10 \end{aligned}$$ (VOELKER, David H., and Orton, Peter Z., 2001)

Substitute n = 1,116 and $p = 0.06$ in the above conditions.

$$\begin{aligned} 1{,}116(0.06) &\approx 67 > 10 \\ 1{,}116(1-0.06) &\approx 1{,}045 > 10 \end{aligned}$$

The conditions are satisfied. The following Minitab output provides the 95% confidence interval for the proportion of golfers earning over $75,000.

CI for One Proportion: Golfer and above $75,000

Event = 1

Variable	X	N	Sample p	95% CI
Golfer and above $75,000	68	1116	0.060932	(0.047623, 0.076610)

Figure 6: 95% Confidence interval for golfers earning over $75,000

From the above Minitab output, the 95% confidence interval for the population proportion of golfers in the U.S earning over $75,000 is (0.05, 0.08). With 95% confidence, it can be estimated that the portion of golfers in the U.S making $75,000 is between 5% and 8%.

Mini case assignment – 2

This assignment aims to verify whether the temperature of hot chocolate at D's Bagels owned by David is maintained at the desired temperature within the desired standard deviation. The data of temperature were collected randomly by David for 25 cups of hot

© 2019, Scientific Research Journal
http://dx.doi.org/10.31364/SCIRJ/v7.i8.2019.P0819XX

chocolate. As the variable temperature is of an interval scale, the appropriate measures of central tendency are mean, median and mode and means of spread are interpretable (WALLER, Derek L., 2008).

The following Minitab output gives the descriptive measures of temperatures of 25 hot chocolate samples.

Descriptive Statistics: Temperature

Variable	Mean	StDev	Q1	Median	Q3	Range	IQR	Mode
Temperature	141.44	1.98	140.00	141.00	143.00	8.00	3.00	140, 141, 143

Variable	N for Mode
Temperature	5

Figure 7: Descriptive statistics – Temperature

Using the Minitab output, it is clear that the sample mean temperature is 141.44 degrees Celsius, the sample median temperature is 143 degrees Celsius, and the mode temperatures are 140, 141, and 143. The sample standard deviation of the heat is 1.98.

The appropriate graphs to display the temperature are histogram and boxplot (R.S.N, Pillai, and V., Bagavathi, 2003). The histogram and boxplot of temperature are as follows.

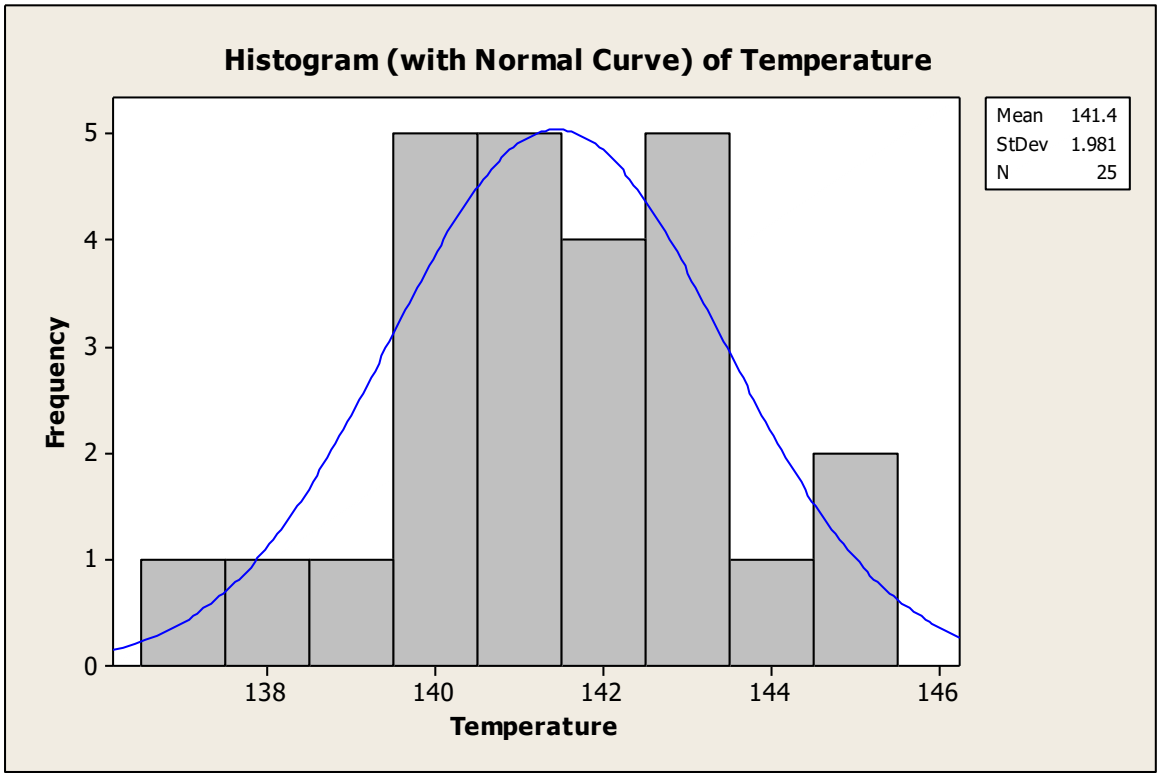

Figure 8: Histogram of temperatures

© 2019, Scientific Research Journal
http://dx.doi.org/10.31364/SCIRJ/v7.i8.2019.P0819XX

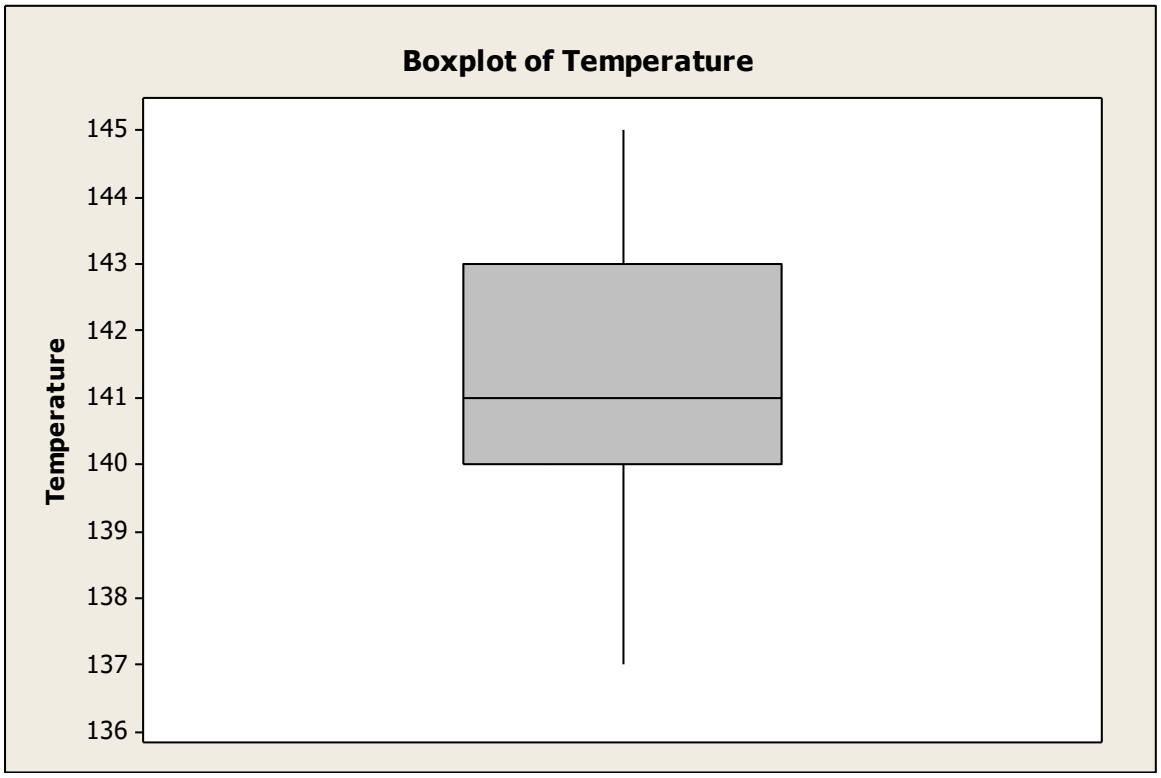

Figure 9: Boxplot of temperatures

The above histogram and boxplot of temperature show that the distribution of temperature is approximately normal, and no outliers are present. To test whether the sample provided sufficient evidence for desired mean temperature and standard deviation of temperature, our hypothesis test offers us the relevant data.

Research hypothesis: Is the sample data evidence to show that the actual mean temperature of the hot chocolate differs from 142 degrees at 10% significance?

Null hypothesis H_0: The correct mean temperature of the hot chocolate does not differ from 142 degrees. $(\mu \neq 142)$.

Alternate hypothesis H_1: The exact mean temperature of the hot chocolate differs from 142 degrees. $(\mu \neq 142)$.

Appropriate test: The proper statistical analysis to verify this claim is the Student's test for the single sample mean as the sample size is less than 30, and the population standard deviation is unknown.

Assumptions:

i) The samples should be random and independent.

ii) The samples should be drawn from the general population. (MONTGOMERY, Douglas C., 2009)

© 2019, Scientific Research Journal
http://dx.doi.org/10.31364/SCIRJ/v7.i8.2019.P0819XX

The validity of assumptions:

i) As the examples are random, randomness and independence are valid.

ii) From the histogram and boxplot, the distribution of temperature is approximately normal.

Test statistic:

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$ (BLUMAN, Allan G, 2009)

Minitab output:

One-Sample T: Temperature

Test of mu = 142 vs not = 142

Variable	N	Mean	StDev	SE Mean	90% CI	T	P
Temperature	25	141.440	1.981	0.396	(140.762, 142.118)	-1.41	0.170

Figure 10: One sample t-test for mean temperature

Decision:

From the above Minitab output, the value of the observed t-statistic value is -1.41 with *p*-value 0.170. As the *p*-value exceeds the level of significance, there is no evidence against the null hypothesis. It is wise to conclude that the correct mean temperature of hot chocolates does not differ from 142 degrees.

Research hypothesis: Is the sample data evidence to show that the correct standard deviation temperature of the hot chocolate is higher than 3 degrees at 10% significance?

Null hypothesis H_0: The true mean, standard deviation temperature of the hot chocolate is less than or equal to 3 degrees. $(\sigma \leq 3)$.

Alternate hypothesis H_1: The accurate mean, standard deviation temperature of the hot chocolate is greater than 3 degrees. $(\sigma > 3)$.

Appropriate test: The most logical statistical analysis to verify this claim is the Chi-square test for a single sample (WALLER, Derek L., 2008).

© 2019, Scientific Research Journal
http://dx.doi.org/10.31364/SCIRJ/v7.i8.2019.P0819XX

Assumptions:

i) The samples should be random and independent.

ii) The samples should be drawn from a healthy population. (MONTGOMERY, Douglas C., 2009)

The validity of assumptions:

i) As the examples are random, randomness and independence are valid.

ii) From the histogram and boxplot, the distribution of temperature is approximately normal.

Test statistic:

$$\sqrt{\frac{(n-1)S^2}{\sigma^2}} \sim \chi^2_{n-1}$$

Minitab output:

Test and CI for One Variance: Temperature

Method

Null hypothesis Sigma = 3
Alternative hypothesis Sigma > 3

Statistics

Variable	N	StDev	Variance
Temperature	25	1.98	3.92

90% One-Sided Confidence Intervals

Variable	Method	Lower Bound for StDev	Lower Bound for Variance
Temperature	Chi-Square	1.68	2.84
	Bonett	1.68	2.83

Tests

Variable	Method	Test Statistic	DF	P-Value
Temperature	Chi-Square	10.46	24	0.992
	Bonett	—	—	1.000

Figure 11: Chi-square test for standard deviation of temperature

Decision:

From the above Minitab output, the value of the observed chi-squared-statistic value is 10.46 with one-sided *p*-value 0.992. As the *p*-value exceeds the level of significance, there is no evidence against the null hypothesis. It is wise to conclude that the actual mean, standard deviation temperature of the hot chocolate is less than or equal to 3 degrees.

© 2019, Scientific Research Journal
http://dx.doi.org/10.31364/SCIRJ/v7.i8.2019.P0819XX

Conclusion

The two mini-case assignments are successfully analyzed using statistical techniques. The result for the first mini-case task shows that the correct proportion of the golfers in the U.S is between 14% and 18% whereas the actual percentage of the golfers in the U.S earning over $75,000 is between 5% and 8%. The result for the second mini-case assignment shows that the correct mean temperature of hot chocolates is 142 degrees, and the actual standard deviation of temperature is less than or equal to 3 degrees. These results are limited to the validity of samples and sampling techniques. It is suggested to get more data samples and compare the results for the reliability of the inferences.

References

1: BLUMAN, Allan, G., 2009. *Elementary statistics A step by step approach, Seventh edition*. New York: The McGraw-Hill Companies Inc.

2: MONTGOMERY, Douglas C., 2009. *Statistical Quality Control A modern introduction, Sixth edition*. New Delhi: Willey India Pvt Ltd.

3: R.S.N, Pillai, and Bagavathi V. 2003. *Statistics*. New Delhi: S.Chand & company.

4: VOELKER, David H., and Peter Z. ORTON. 2001. *Cliffs quick review of Statistics*. New York: Hungry Minds, Inc.

5: WALLER, Derek L., 2008. *Statistics for Business*. Burlington: Elsevier, Inc.

6: ZIKMUND, William G., 2003. *Business research methods, Seventh edition*. Cincinnati: Thomson / South-Western.

© 2019, Scientific Research Journal
http://dx.doi.org/10.31364/SCIRJ/v7.i8.2019.P0819XX