MIS771 Descriptive Analytics and Visualisation

Topic 2 Tutorial – Hypothesis Testing and comparing two samples

BRIEF ANSWERS

Q1. Time Spent per visit

300 randomly selected customers were participated in the customer research survey and it is found that, on average, a BLITZ customer spends 40.44 minutes during a visit to any of the BLITZ stores with a standard deviation of 11.46 minutes. ($\bar{x}$ = 40.44, s = 11.46)

The corresponding figure for average time spent per customer visit last year was 38 minutes. Management would like to know if this year's average time spent is more than the last year's average.

(a) Explain why they should use hypothesis testing in this situation?

Testing a Claim

(b) What type of variable are we working with?

Numerical Variable

(c) Is this a one sample or two sample situation?

One

(d) Use Excel to complete the hypothesis test (use α=1%). Interpret your results.

Hypothesis Test for μ			
Hypotheses			
Null Hypothesis	μ	0	**38**
Alternative Hypothesis	μ	>	**38**
Test Type			Upper
Level of significance			
		α	**0.01**
Critical Region			
Degrees of Freedom			299
Critical Value			2.3389
Sample Data			
Sample Standard Deviation			**11.46**
Sample Mean			**40.44**
Sample Size			**300**
Standard Error of the Mean			0.6616

t Sample Statistic	3.6878
p-value	0.0001
Decision	
Reject Null Hypothesis	

At 1 percent significance level, there is sufficient evidence to conclude that this year's average time spent shopping for all BLITZ customers is more than the last year's average of 38 minutes.

Q2. BLTIZ Online Store

Out of the 300 customers participated in the survey it is found that only 66 members have visited BLITZ online store. BLITZ have set a KPI to have at least 25% of their customers who visit their stores to also visit their online store. They are very much interested to know whether the survey result is significantly lower than their expectation.

(a) Explain why we can use hypothesis testing in this situation.

Comparing against a standard.

(b) What type of variable are we working with?

Categorical Variable.

(c) Is this a one sample or two sample situation?

One

(d) Perform a hypothesis test in Excel. (Use α=5%) What p-value did you obtain? Interpret this value.

Hypothesis Test for π			
Hypotheses			
Null Hypothesis	π	0	**25%**
Alternative Hypothesis	π	<	**25%**
Test Type			Lower
Level of significance			
		α	**0.05**
Critical Region			
Critical Value			-1.6449
Sample Data			
Sample Size			**300**
Count of 'Successes'			**66**
Sample proportion, p			**22.00%**
Standard Error			**2.50%**
z Sample Statistic			-1.2000
p-value			0.1151
Decision			
Fail to reject Null Hypothesis			

At 5 percent significance level, there is not enough evidence to conclude that less than 25% of all customers who visit BLITZ stores also visit their online store.

p-value = 11.5%

There would be a 11.5% chance of making a Type I error (that is if we were to conclude less than 25% of all customers who visit their stores also visit their online store when in fact it is equal to or above 25%).

BLITZ customers – Time spent shopping across cities

Following on from Q1, given we now know that the time spent shopping is more than last year, we wish to estimate *how much* more time customers are spending shopping.

(a) Calculate a 95% confidence interval for the average time spent shopping for all customers?
We are 95 percent confident that the true average time spent shopping for all BLITZ customers is between 39.09 to 41.71 minutes.

(b) Calculate a 90% confidence interval for the average time spent shopping for all customers?
39.30 to 41.50

(c) Calculate a 99% confidence interval for the average time spent shopping for all customers?
38.68 to 42.12

(d) Comment on the differences between the three confidence intervals.
The higher the confidence the less accurate an answer (wider interval).

BLITZ management would like to know whether the time spent differs across cities. Particularly they would like to know whether there is *a difference* in average time in Melbourne and Perth customers.

	Melbourne	**Perth**
$\bar{x}$	46.5	30.0
s	11.2	4.5
n	100	100

(a) Perform a hypothesis test using Excel to clarify BLITZ management on this. (Use α= 5%)

Hypotheses			
Null Hypothesis	$\mu_1 - \mu_2$	=	0
Alternative Hypothesis	$\mu_1 - \mu_2$	≠	
Test Type			Two
Level of significance			
		α	0.05
Critical Region			
Degrees of Freedom			130
Lower Critical Value			-1.9784
Upper Critical Value			1.9784
Sample Results			
Sample 1 Data			
Sample Standard Deviation			11.20
Sample Mean			46.50
Sample Size			100
Sample 2 Data			
Sample Standard Deviation			4.50
Sample Mean			30.00
Sample Size			100
Standard Error of the Mean			1.2070
t Sample Statistic			13.6700
p-value			0.0000
Decision			
Reject Null Hypothesis			

(b) Based on the computer output created above, briefly advise BLITZ management about the findings.

There is sufficient evidence to conclude that there is a difference in the average time spent shopping between Melbourne and Perth customers. Melbourne customers on average shop for longer.

Note: Would be useful to do a CI as well to estimate the size of the difference.

Q4. Loyalty program awareness across different age groups

BLITZ management also would like to know whether younger customers are more likely to be aware of the loyalty program compared to older customers. Out of the 71 customers aged under 30 years old, 31 were aware of the program. Of the 229 customers aged 30 or above, 83 were aware of the program.

(a) Explain why we can use hypothesis testing in this situation?

Testing a claim

(b) Is this a one sample or two sample situation?

Two

(c) Write down H0 and H1 that are appropriate for this problem. (Write them down in words and in symbols.)

H_0: $\pi_1 - \pi_2 \leq 0\%$

Younger customers are the same or less likely to be aware of the loyalty program compared to older customers

H_1: $\pi_1 - \pi_2 > 0\%$

Younger customers are more likely to be aware of the loyalty program compared to older customers

(d) Use Excel to complete the hypothesis test (use α=5%).

Hypothesis Test for π1 - π2			
Hypotheses			
Null Hypothesis	$\pi_1 - \pi_2$	0	**0%**
Alternative Hypothesis	$\pi_1 - \pi_2$	>	**0%**
Test Type			Upper
Level of significance			
		α	**0.05**
Critical Region			
Critical Value			1.6449
Sample Data			
Sample 1 Data			
Sample Size			**71**
Count of 'Successes'			**31**
Sample proportion, p_1			**43.66%**
Sample 2 Data			
Sample Size			**229**
Count of 'Successes'			**83**
Sample proportion, p_2			**36.24%**
Pooled estimate of proportion			**38.00%**
Standard Error			**6.59%**
z Sample Statistic			1.1250
p-value			0.1303
Decision			
Fail to reject Null Hypothesis			

(e) Would it be different if we use α=1% or α=10%?

No p-value = 13% > α

(f) Interpret the output to BLITZ management?

At 5 percent significance, there is not enough evidence to conclude that younger customers are more likely to be aware of the loyalty program compared to older customers.

Q5. Average spend across different loyalty tiers

The membership structure is set up in a way to reward members more according to their level of membership. To justify the extra costs associated with higher levels of membership, management want an estimate in the difference spent between levels (in particular between BLUE tier members and SILVER tier members).

This year's survey included 28 BLUE tier members and 24 SILVER tier members. Based on the survey results it is found out that BLUE tier average spend per visit is $97.32 with a standard deviation of $24.55 and SILVER tier average spend per visit is $109 with a standard deviation of $27.05.

(a) Explain why a confidence interval is appropriate in this case.

Estimate

(b) What assumptions do you have to make about the data?

As n < 30 need to assume data is approx. normal

(c) Calculate a 95% confidence interval and interpret your answer.

Confidence Interval for μ1 - μ2 (independent, equal variance	
Level of Confidence	
Level of Confidence	**95%**
Sample Results	
Sample 1 Data	
Sample Standard Deviation	**24.55**
Sample Mean	**97.32**
Sample Size	**28**
Sample 2 Data	
Sample Standard Deviation	**27.05**
Sample Mean	**109**
Sample Size	**24**
Intermediate Calculations	
Degrees of Freedom	50
Pooled Variance	662.04
Standard Error of the Mean	7.1575
t value	2.0086
Confidence Interval for μ1 - μ2	
Interval Lower Limit	**-26.06**
Interval Upper Limit	**2.70**

As the CI contains both negative and positive values it is inconclusive. Unable to say whether one mean is different to the other.