LAB -9

Two sample Sample Z-test

Challenging Experiment -6

Aim:-To test the hypothesis for Large Samples by using Two–sample Z-test

R-Code:-

Tests about a proportion using x and n using the prop.test function:

Usage: prop.test(c(x1,x2), c(n1,n2), correct=, alternate =).

1. x1 and x2 are the number of successes in sample 1 and 2 respectively.
2. n1 and n2 are the sample sizes or number of trials.
3. correct = TRUE (use a continuity correction factor) or FALSE (do not).
4. alternate = "two.sided" (default), "less", or "greater".

Problem : A popular cold-remedy was tested for it's efficacy. In a sample of 150 people who took the remedy upon getting a cold, 117 (78%) had no symptoms one week later. In a sample of 125 people who took the placebo upon getting a cold, 90 (75%) had no symptoms one week later. The table summarizes this information.

Group	***#who are symptom Free after one week(x)***	***Total # in group (n)***	***Proportion*** $\hat{p} = x / n$
Remedy	***117***	***150***	***0.78***
Placebo	***90***	***120***	***0.75***

The Test: Test the claim that the proportion of all remedy users who are symptom-free after one week is greater than the proportion for placebo users. Test this claim at the 0.05 significance level.

R Code:-

```
> x<-c(117,90)
> n<-c(150,120)
> prop.test(x,n,alternative="greater",correct=FALSE)

        2-sample test for equality of proportions without continuity
        correction

data:  x out of n
X-squared = 0.3354, df = 1, p-value = 0.2812
alternative hypothesis: greater
95 percent confidence interval:
 -0.05557192  1.00000000
sample estimates:
prop 1 prop 2 
  0.78   0.75 
```

We fail reject the null hypothesis because the P-value (.2812) is greater than the significance level. Therefore, we can't support the claim.

Problem 2:-

The Trial Urban District Assessment (TUDA) is a study sponsored by the government of student achievement in large urban school district. In 2009, 1311 of a random sample of 1900 eighth-graders from Houston performed at or above the basic level in mathematics . In 2011, 1440 of a random sample of 2000 eighth-graders from Houston performed at or above the basic level . (The study reports the proportions).

(A)Is there an increase in the proportion of eighth-graders who performed at or above the basic level in mathematics from 2009 to 2011 at the 5% significance level?

(B) Compute the 95% confidence interval for the difference in proportion of eighth-graders who performed at or above the basic level in mathematics from 2009 to 2011.

Solution:-

Let p1 and p2 be the proprtions of eighth-graders that performed at or above he basic level in mathematics in 2011 and 2009,respectively .We wish the test

H_0=p1=p2 against H_1=p1> p2

```
> prop.test(c(1440,1311),c(2000,1900),alternative="greater",correct=FALSE)

        2-sample test for equality of proportions without continuity
        correction

data:  c(1440, 1311) out of c(2000, 1900)
X-squared = 4.2197, df = 1, p-value = 0.01998
alternative hypothesis: greater
95 percent confidence interval:
 0.005972807 1.000000000
sample estimates:
prop 1 prop 2
  0.72   0.69
```

The p-value=0.02 < 0.05 so we reject H0. Thus, there is evidence that there is an increase from 2009 to 2011 in the proportion of eighth-graders who performed at or above the basic level at the 5% significance level.

Solution to part (b). We obtain

```
> prop.test(c(1440,1311),c(2000,1900),correct=FALSE)

        2-sample test for equality of proportions without continuity
        correction

data:  c(1440, 1311) out of c(2000, 1900)
X-squared = 4.2197, df = 1, p-value = 0.03996
alternative hypothesis: two.sided
95 percent confidence interval:
 0.001369833 0.058630167
sample estimates:
prop 1 prop 2
  0.72   0.69
```

Thus, we are 95% confident that the percent of eighth-graders who performed at or above the basic level in mathematics in 2011 is between 0:14% and 5:86% higher than in 2009.

Problem. The use of helmet among recreational alpine skiers and snowboarders are generally low. A study from Norway wanted to examine if helmet use reduces the risk of head injury. In the study, they compared the helmet use among skiers and snowboarders that was injured with a control group. The control group consisted of skiers and snowboarders that was uninjured. 96 of 578 people with head injuries used a helmet and 656 of 2992 people in the uninjured group used a helmet. Is helmet use lower among skiers and snowboarders who had head injuries?

Solution : Let p1 be the proportion of helmet use among injured skiers and snowboarders. Let p2 be the proportion of helmet use among uninjured skiers and snowboarders.

We wish to test

H0 : p1 = p2 against Ha : p1 < p2

```
> prop.test(c(96,656),c(578,2992),alternative=c("less"),correct=FALSE)

        2-sample test for equality of proportions without continuity
        correction

data:  c(96, 656) out of c(578, 2992)
X-squared = 8.2336, df = 1, p-value = 0.002056
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000 -0.02482216
sample estimates:
   prop 1    prop 2
0.1660900 0.2192513
```

The p-value= 0.0021 < 0.01 so we have strong evidence that helmet use is lower among skiers and snowboarders who had head injuries compared to uninjured skiers and snowboarders.

Problem : A survey is taken two times over the course of two weeks. The pollsters wish to see if there is a difference in the results as there has been a new advertising campaign run. Here is the data

	Week1	*Week2*
Favorable	*45*	*56*
Unfavorable	*35*	*47*

The standard hypothesis test is H0: P1 = P2 against the alternative (two-sided) H1: P1 =P2. The function prop.test is used to being called as prop.test(x,n) where x is the number favorable and n is the total. Here it is no different, but since there are two x's it looks slightly different. Here is how

```
> prop.test(c(45,56),c(45+35,56+47))

        2-sample test for equality of proportions with continuity correction

data:  c(45, 56) out of c(45 + 35, 56 + 47)
X-squared = 0.010813, df = 1, p-value = 0.9172
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1374478  0.1750692
sample estimates:
   prop 1    prop 2
0.5625000 0.5436893
```

we observe that the p-value is 0.9172 so we accept the null hypothesis that P1 =P2.

Problem

In the built-in data set named quine, children from an Australian town is classified by ethnic background, gender, age, learning status and the number of days absent from school.

> library (MASS)   # load the MASS package

> quine

> table (quine$Eth, quine$Sex)

	F	M
A	38	31
N	42	35

Assuming that the data in quine follows the normal distribution, find the 95% confidence interval estimate of the difference between the female proportion of Aboriginal students and the female proportion of Non-Aboriginal students, each within their own ethnic group.

Ans)

Solution

We apply the prop.test function to compute the difference in female proportions. The Yates's continuity correction is disabled for pedagogical reasons.

```
> prop.test(table(quine$Eth, quine$Sex), correct=FALSE)

        2-sample test for equality of proportions without continuity
        correction

data:  table(quine$Eth, quine$Sex)
X-squared = 0.0040803, df = 1, p-value = 0.9491
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1564218  0.1669620
sample estimates:
   prop 1    prop 2
0.5507246 0.5454545
```

.

Conclusion:

The 95% confidence interval estimate of the difference between the female proportion of Aboriginal students and the female proportion of Non-Aboriginal students is between -15.6% and 16.7%.

Comparison of two independent sample means, taken from two populations with known variance

Problem: The following data shows the heights of individuals of two different countries with the population variance of 5 and 8.5 respectively. Is there any significant difference between the average heights of two groups.

A: 175	168	168	190	156	181	182	175	174	179
B: 185	169	173	173	188	186	175	174	179	180

Since we have the variance of the population, we must proceed with a two sample Z-test.

Solution :

```
z.test2sam = function(a, b, var.a, var.b)

{

n.a = length(a)

n.b = length(b)

zeta = (mean(a) - mean(b)) / (sqrt(var.a/n.a + var.b/n.b))

return(zeta)

}

a = c(175, 168, 168, 190, 156, 181, 182, 175, 174, 179)

b = c(185, 169, 173, 173, 188, 186, 175, 174, 179, 180)

z.test2sam(a, b, 5, 8.5)
```

Inference :-

The value of z is greater than the critical value of z tabulated for alpha equal to 0.05 (z-tabulated = 1.96 for a two-tailed test): we reject the null hypothesis and conclude that the two means are significantly different.

Practice problems :

1. In the large city A,20 per cent of Random sample of 900 School children had defective eye –sight. In the large city B,15 percent of random sample of 1600 school children had the same defective. Is this Difference between the two Proportions Significant? Obtain 95% confidence limits of the difference in the population proportions.

2. A cigarette manufacturing firm claims its brand A of the cigarettes outsells its brand B by 8%.if its found that 42 out sample of 200 smoker prefer brand A and 18 out of another random sample of 100 smokers prefers brand B, test whether the 8% difference is a valid cliam.

Note:- In real time ,even though sample size is greater than 30 we should use t-test for means .