MSIS 5303 – Statistics for Data Science – Fall 2021 - Assignment 8 Solution

1) (**1 point**) Parents of teenage boys often complain that auto insurance costs more, on average, for teenage boys than for teenage girls. A group of concerned parents examines a random sample of insurance bills. The mean annual cost for 36 teenage boys was $679. For 23 teenage girls, it was $559. From past years, it is known that the population standard deviation for each group is $180. Determine whether or not you believe that the mean cost for auto insurance for teenage boys is greater than that for teenage girls.

   Independent samples. We assume normal populations.

H_0: $\mu_1 - \mu_2 = 0$ (mean auto insurance costs for teenage boys = mean auto insurance costs for teenage girls)
H_a: $\mu_1 - \mu_2 > 0$ ((mean auto insurance costs for teenage boys > mean auto insurance costs for teenage girls)

```
> n1 = 36
> n2 = 23
> x1_bar = 679
> x2_bar = 559
> sigma1 = 180
> sigma2 = 180
> se1 <- (sigma1^2/n1)
> se2 <- (sigma2^2/n2)
> std_err <- sqrt(se1+ se2)
> print(std_err)
[1] 48.04889
> z_statistic <- (x1_bar - x2_bar)/std_err
> print(z_statistic)
[1] 2.497456
> # Lower-tailed test
> #p_value <- (pnorm(z_statistic,deg_free))
> #Upper-tailed test
> p_value <- (1 - pnorm(z_statistic,0,1))
> # Two-tailed test
> #p_value <- 2*(1-*pnorm(abs(z_statistic,0,1)))
> print(p_value)
[1] 0.006254394
```

Since p =0.006 < α = 0.05, **we reject the null hypothesis** and conclude that the mean cost for auto insurance for teenage boys is significantly greater than that for teenage girls at α = 0.05.

2) (**2 points**) The Eastern and Western Major League Soccer conferences have a new Reserve Division that allows new players to develop their skills. Data for a randomly picked date showed the following annual goals.

Western	Eastern
Los Angeles 9	D.C. United 9
FC Dallas 3	Chicago 8
Chivas USA 4	Columbus 7
Real Salt Lake 3	New England 6
Colorado 4	MetroStars 5
San Jose 4	Kansas City 3

Is there sufficient evidence to conclude that the **W** Division teams score more goals, on average, than the **E** teams at $\alpha = 0.05$? *Do the problem using both the t-test() function in R as well as using the detailed code for performing the test and compare the results.*

Two independent samples. The population standard deviations are unknown so we use the Welch Two-sample t-test.
H_0: $\mu_1 - \mu_2 = 0$ (mean number of goals costs for **W** teams = mean number of goals for **E** teams)
H_a: $\mu_1 - \mu_2 > 0$ (mean number of goals costs for **W** teams > mean number of goals for **E** teams)

```
> western <- c(9,3,4,3,4,4)
> eastern <- c(9,8,7,6,5,3)
> t.test(western,eastern,paired=FALSE, alternative = "greater")


        Welch Two Sample t-test

data:  western and eastern
t = -1.437, df = 9.9804, p-value = 0.9093
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -4.146213       Inf
sample estimates:
mean of x mean of y
 4.500000  6.333333


> n1 = length(western)
> n2 = length(eastern)
> x1_bar = mean(western)
> x2_bar = mean(eastern)
> s1 = sd(western)
> s2 = sd(eastern)
> se1 <- (s1^2/n1)
> se2 <- (s2^2/n2)
> std_err <- sqrt(se1+ se2)
> print(std_err)
[1] 1.275844
> t_statistic <- (x1_bar - x2_bar)/std_err
> print(t_statistic)
[1] -1.436957
> deg_free_n <- (se1 + se2)^2
> deg_free_d <- (se1^2/(n1 -1)) + (se2^2/(n2 -1))
> deg_free <- deg_free_n/deg_free_d
> print(deg_free)
[1] 9.980353
> # Lower-tailed test
> p_value <- 1 - (pt(t_statistic,deg_free))
> print(p_value)
[1] 0.9093349
```

Since $p = 0.909 > \alpha = 0.05$, **we fail reject the null hypothesis** and conclude that there is no significant difference in the mean number of goals for **W** and **E** teams at $\alpha = 0.05$.

3) (**1 point**) Joan Nguyen recently claimed that the proportion of college-age males with at least one pierced ear is as high as the proportion of college-age females. She conducted a survey in her classes. Out of 107 males, 20 had at least one pierced ear. Out of 92 females, 47 had at least one pierced ear. Test using $\alpha = 0.01$ whether the proportion of males (with at least one ear pierced) is the same as the proportion of females.

Independent samples Test of difference in proportions.

H_0: $p_1 - p_2 = 0$ (proportion of college-age males with at least one pierced ear = proportion of college-age females with at least one pierced ear)
H_a: $p_1 - p_2 \neq 0$ (proportion of college-age males with at least one pierced ear $\neq$ proportion of college-age females with at least one pierced ear)

```
> xA = 47
> nA = 92
> xB = 20
> nB = 107
> pprime_A = xA/nA
> pprime_B = xB/nB
> pc = (xA + xB)/(nA + nB)
> std_err <- sqrt(pc*(1-pc)*((1/nA)+(1/nB)))
> print(std_err)
[1] 0.06719114
> z_statistic <- (pprime_A - pprime_B)/std_err
> print(z_statistic)
[1] 4.821375
> # two-tailed test
> p_value <- 2*(1-pnorm(z_statistic,0,1))
> print(p_value)
[1] 1.425722e-06
```

Since p =0.000001 < α = 0.01, **we reject the null hypothesis** and conclude that the proportion of college-age males with at least one pierced ear is significantly different from proportion of college-age females with at least one pierced ear at α = 0.01.

4) (**1 point**) A traveler wanted to know if the prices of hotels are different in the ten cities that he visits the most often. The list of the cities with the corresponding hotel prices for his two favorite hotel chains is in the Table. Test at the 1% level of significance whether there is a difference in prices. *Do the problem using both the t-test() function in R as well as using the detailed code for performing the test and compare the results.*

Cities	Hyatt Regency prices in dollars	Hilton prices in dollars
Atlanta	107	169
Boston	358	289
Chicago	209	299
Dallas	209	198
Denver	167	169
Indianapolis	179	214
Los Angeles	179	169
New York City	625	459
Philadelphia	179	159
Washington, DC	245	239

This is a test of dependent samples or match pairs.

Let μ_d be the mean difference in prices between the two hotel chains.

H_0: $\mu_d = 0$
Ha: $\mu_d \neq 0$

```
> Hyatt <- c(107, 358, 209, 209, 167, 179, 179, 625, 179, 245)
> Hilton <- c(169, 289, 299, 198, 169, 214, 169, 459, 159, 239)
> xd <- Hyatt - Hilton
> x_bar_d <-mean(xd)
> std_err <- sd(xd)/sqrt(length(xd))
> print(x_bar_d)
[1] 9.3
> print(std_err)
[1] 22.43066
> t_statistic <- x_bar_d/std_err
> print(t_statistic)
[1] 0.4146111
> deg_free <- length(xd) - 1
> print(deg_free)
[1] 9
> # Lower-tailed test
> #p_value <- (pt(t_statistic,deg_free))
> # Upper-tailed test
> # p_value <- (1 - pt(t_statistic,deg_free))
> # Two-tailed test
> p_value <- abs(2*(1-pt(t_statistic,deg_free)))
> print(p_value)
[1] 0.6881346
> #
```

```
> t.test(Hyatt,Hilton,paired=TRUE, alternative = "two.sided", conf.level = 0.99)

        Paired t-test

data:  Hyatt and Hilton
t = 0.41461, df = 9, p-value = 0.6881
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -63.59595  82.19595
sample estimates:
mean of the differences
                    9.3
```

Since $p = 0.6881 > \alpha = 0.01$, **we fail reject the null hypothesis** and conclude that there is no significant difference in mean prices between the two hotel chains at $\alpha = 0.01$.