NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

R Lesson 10 – Solutions
MSPA 401 – Introduction to Statistics

1) A double-blind clinical trial of a new drug for back pain was designed using control and treatment groups. Volunteers were fully informed and assigned at random to each group. Neither the volunteers nor the doctor knew when the new drug or a placebo was being administered. When 100 volunteers in each group had been treated and evaluated, the results revealed an 85% success rate for the new drug and a 65% success rate for the control group. At the 95% confidence level, is there a statistically significant difference between the two reported rates? Use a one-sided test. Report a confidence interval for the difference.

```
         success fail
new_drug      85   15
control       65   35
> prop.test(x, alternative = "greater", conf.level = 0.95)

 2-sample test for equality of proportions with continuity correction

data:  x
X-squared = 9.6267, df = 1, p-value = 0.0009589
alternative hypothesis: greater
95 percent confidence interval:
 0.09199653 1.00000000
sample estimates:
prop 1 prop 2
  0.85   0.65
```

p-value = 0.0009589 < 0.05 (reject null hypothesis)

2) Two baseball players had their career records compared. In 267 times at bat, one player hit 85 home runs. In 248 times at bat, the other player hit 89 home runs. Assume the number of home runs follows a binomial distribution, is there a statistically significant difference with 95% confidence between the home run averages for these two baseball players?

```
         HR Other
Player A 85   182
Player B 89   159
> prop.test(x, alternative = "two.sided", conf.level = 0.95)

 2-sample test for equality of proportions with continuity correction

data:  x
X-squared = 0.7712, df = 1, p-value = 0.3799
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.12617606  0.04513824
sample estimates:
   prop 1    prop 2
0.3183521 0.3588710
```

p-value = 0.3799 > 0.05 (fail to reject null hypothesis)

3) Using the **home_prices.csv** data, compare mean selling prices between homes located in the northeast sector of the city versus the remaining homes. Also, compare the mean selling prices

between homes with a corner lot and those located elsewhere. Use two-sample t-tests for the hypothesis tests at the 95% confidence level. Report confidence intervals for each.

```
> with(homes, by(PRICE, NBR, summary))  # price stats across sectors
NBR: NO
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1350    1920    2350    2458    2625    5250
------------------------------------------------------------
NBR: YES
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1548    2016    2462    2732    3125    5375
> with(homes, by(PRICE, CORNER, summary))  # price stats across corner or not
CORNER: NO
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1350    1974    2388    2657    3044    5375
------------------------------------------------------------
CORNER: YES
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1748    1939    2469    2571    2829    5250
```

```
        Welch Two Sample t-test

data:  NE_PRICE and OTHER_PRICE
t = 1.6, df = 83.277, p-value = 0.1134
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -66.56374 614.32015
sample estimates:
mean of x mean of y
 2731.891  2458.013
```

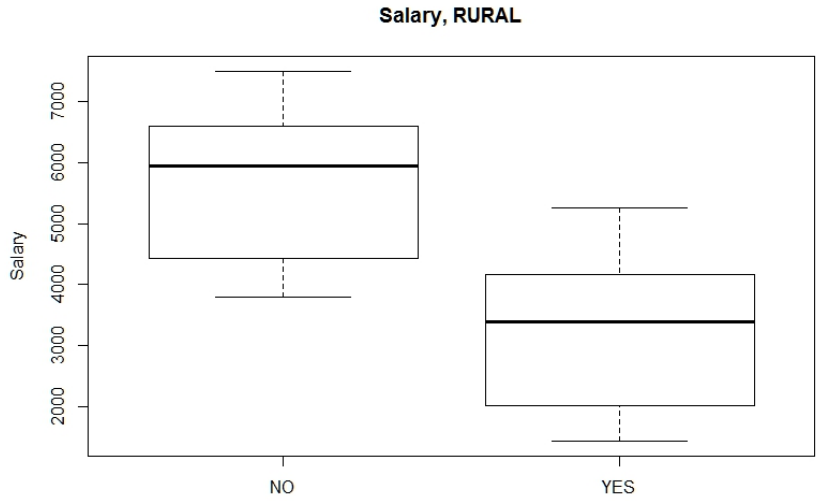
p-value = 0.1134 > 0.05 (fail to reject null hypothesis; prices of homes in the NE are not statistically different from prices of other homes).

```
        Welch Two Sample t-test

data:  CORNER_PRICE and NON_CORNER_PRICE
t = -0.4319, df = 34.664, p-value = 0.6685
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -490.9729  318.7576
sample estimates:
mean of x mean of y
 2570.682  2656.789
```

p-value = 0.6685 > 0.05 (do not reject null hypothesis; prices of homes on corners are not statistically different from non-corner homes).

4) The **nsalary.csv** data are derived from data collected by the Department of Social Services of the State of New Mexico. The data have been adapted for this problem. Present a boxplot comparing RURAL and non-RURAL salaries. Using these data compare mean salary levels between RURAL and non-RURAL locations. Use a two-sample t-test at the 95% confidence level. Report your results.




```
> with(nsalary, by(NSAL, RURAL, summary))  # price stats across sectors
RURAL: NO
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3803    4432    5946    5670    6590    7489
------------------------------------------------------------
RURAL: YES
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1445    2040    3386    3251    4158    5257


        Welch Two Sample t-test

data:  RURAL_SALARY and NON_RURAL_SALARY
t = -5.8555, df = 20.812, p-value = 8.504e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3277.645 -1558.961
sample estimates:
mean of x mean of y
 3251.312  5669.615
```

p-value = 8.504e-06 < 0.05 (reject null hypothesis, there are statistically significant differences between rural and non-rural salaries).

5) **tires.csv** contains data published by R.D. Stichler, G.G. Richey, and J. Mandel, "Measurement of Treadware of Commercial Tires, Rubber Age, 73:2 (May 1953). Treadwear measures of tires each tire was subject to measurement by two methods, the first based on weight loss and the second based on groove wear. Use a paired t-test at the 95% confidence level to test for a difference between the two methods. Report your results using a confidence interval.


```
        Paired t-test

data:  WGT and GRO
t = 5.6503, df = 15, p-value = 4.614e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.837493 6.275007
sample estimates:
mean of the differences
                4.55625
```

p-value = 4.614e-05 < 0.05 (reject the null hypothesis that the means of the two measures are identical. There are statistically significant differences between these two measures of tire wear).