Lesson 10: Statistical Inferences About Two Populations

References

- Black, Chapter 10 Statistical Inferences About Two Populations (pp. 316-358)
- Kabakoff, Chapter 7 Basic Statistics (pp. 158-160)
- Davies, Chapter 18 Hypothesis Testing (pp. 384-433)
- Stowell, Chapter 6 Tabular Data (pp. 73-86), Chapter 10 Hypothesis Testing (pp. 144-146, 158)

Exercises:

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. Also, report a confidence interval for the difference.

```
x <- matrix(c(85,65,15,35), nrow = 2, ncol = 2, byrow = FALSE,
    dimnames = list(c("new_drug", "control"),c("success", "fail")))
print(x)
```

```
##          success fail
## new_drug      85   15
## control       65   35
```

```
prop.test(x, alternative = "greater", conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  x
## X-squared = 10.667, df = 1, p-value = 0.0005454
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.1019965 1.0000000
## 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?

```
# (First, please note that these players had short careers in a league
# somewhere on a planet with a weak gravitational force.)
x <- matrix(c(85,89,(267-85),(248-89)), nrow = 2, ncol = 2, byrow = FALSE,
    dimnames = list(c("Player A", "Player B"), c("HR", "Other")))
print(x)
```

```
##          HR Other
## Player A 85   182
## Player B 89   159
```

```r
prop.test(x, alternative = "two.sided", conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  x
## X-squared = 0.94359, df = 1, p-value = 0.3314
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.12228727  0.04124945
## sample estimates:
##    prop 1    prop 2
## 0.3183521 0.3588710
```

```r
# p-value = 0.3799 > 0.05 (do not reject null hypothesis, the difference
# between the home run rates of these players is nonsignificant.)
```

3) Using the home_prices.csv data (described in Lesson 1), 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.

```r
# Read the comma-delimited text file creating a data frame object in R,
# then examine its structure:

houses <- read.csv("home_prices.csv")
str(houses)
```

```
## 'data.frame':    117 obs. of  8 variables:
##  $ PRICE : num  1350 2550 1550 1828 1800 ...
##  $ SQFT  : int  1142 1478 1480 1299 1121 1400 1505 1050 900 1215 ...
##  $ YEAR  : int  1959 1961 1965 1967 1968 1969 1969 1970 1971 1971 ...
##  $ BATHS : num  1.5 2 1.5 1 1.5 1.5 1.5 1 1 1.5 ...
##  $ FEATS : int  0 3 4 6 4 1 2 1 3 3 ...
##  $ NBR   : Factor w/ 2 levels "NO","YES": 1 2 1 2 2 1 1 2 1 2 ...
##  $ CORNER: Factor w/ 2 levels "NO","YES": 1 2 1 1 1 2 2 1 1 1 ...
##  $ TAX   : num  558 1565 1275 1462 995 ...
```

```r
# Assume that NBR = YES is for the northeast sector of the city
print(summary(houses))  # overall descriptive statistics
```

```
##      PRICE           SQFT           YEAR           BATHS
##  Min.   :1350   Min.   : 837   Min.   :1959   Min.   :1.000
##  1st Qu.:1950   1st Qu.:1280   1st Qu.:1991   1st Qu.:1.000
##  Median :2400   Median :1549   Median :1999   Median :1.500
##  Mean   :2641   Mean   :1645   Mean   :1996   Mean   :1.585
##  3rd Qu.:3000   3rd Qu.:1894   3rd Qu.:2008   3rd Qu.:2.000
##  Max.   :5375   Max.   :2931   Max.   :2013   Max.   :3.000
##      FEATS        NBR      CORNER        TAX
##  Min.   :0.00   NO :39   NO :95   Min.   : 557.5
##  1st Qu.:3.00   YES:78   YES:22   1st Qu.:1477.5
##  Median :4.00                     Median :1807.5
```

```
##  Mean   :3.53                         Mean   :1989.8
##  3rd Qu.:4.00                         3rd Qu.:2307.5
##  Max.   :8.00                         Max.   :4412.5
```

```r
with(houses, 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
```

```r
with(houses, 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
```

```r
# Now, we are ready to do the hypothesis tests.
NE_PRICE <- subset(houses, subset = (NBR == "YES"))$PRICE
OTHER_PRICE <- subset(houses, subset = (NBR == "NO"))$PRICE
t.test(NE_PRICE, OTHER_PRICE, alternative = "two.sided", conf.int = 0.95)
```

```
##
##  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
```

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

CORNER_PRICE <- subset(houses, subset = (CORNER == "YES"))$PRICE
NON_CORNER_PRICE <- subset(houses, subset = (CORNER == "NO"))$PRICE
t.test(CORNER_PRICE, NON_CORNER_PRICE, alternative = "two.sided", conf.int = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  CORNER_PRICE and NON_CORNER_PRICE
## t = -0.43192, 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. 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.

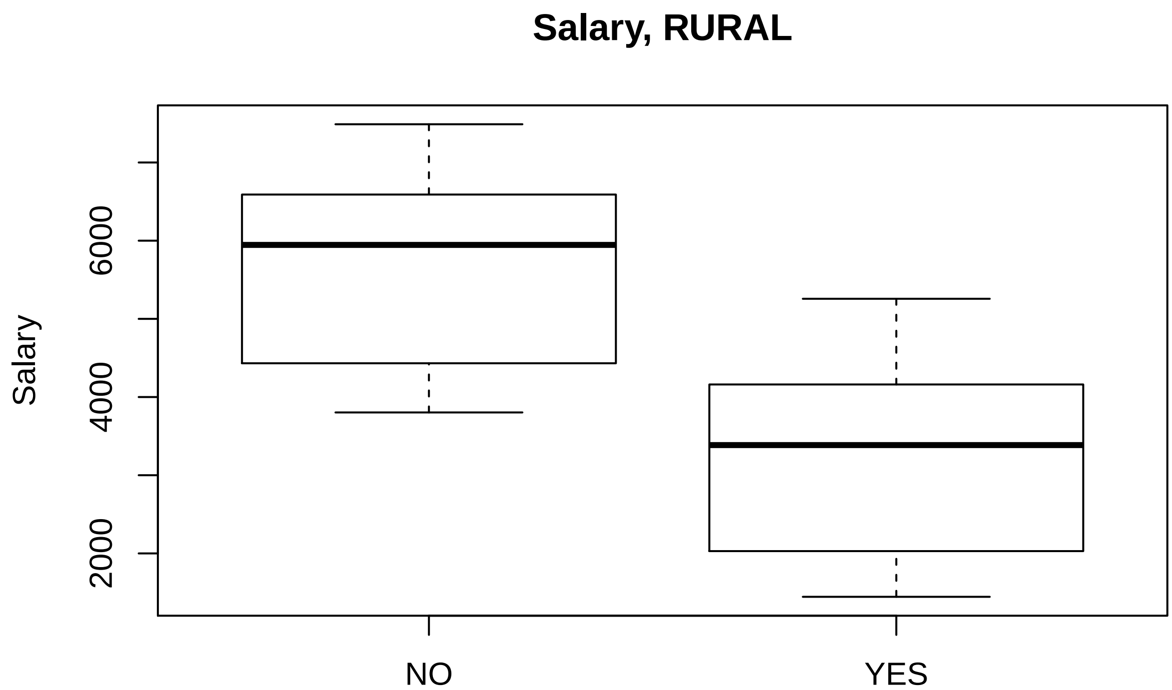
```
# Read the comma-delimited text file creating a data frame object in R,
# then examine its structure:

nsalary <- read.csv("nsalary.csv")
str(nsalary)
```

```
## 'data.frame':    45 obs. of  2 variables:
##  $ RURAL: Factor w/ 2 levels "NO","YES": 2 1 1 1 2 2 2 2 1 2 ...
##  $ NSAL : int  2459 6304 6590 5362 3622 4406 4173 3224 5946 1925 ...
```

```
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
```

```
# Create comparative boxplot
with(nsalary, boxplot(NSAL ~ RURAL, main = "Salary, RURAL",
    ylab = "Salary"))
```




```r
# salaries obviously different for rural vs. non-rural

RURAL_SALARY <- subset(nsalary, subset = (RURAL == "YES"))$NSAL
NON_RURAL_SALARY <- subset(nsalary, subset = (RURAL == "NO"))$NSAL
t.test(RURAL_SALARY, NON_RURAL_SALARY, alternative = "two.sided", conf.int = 0.95)
```

```
##
##  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
```

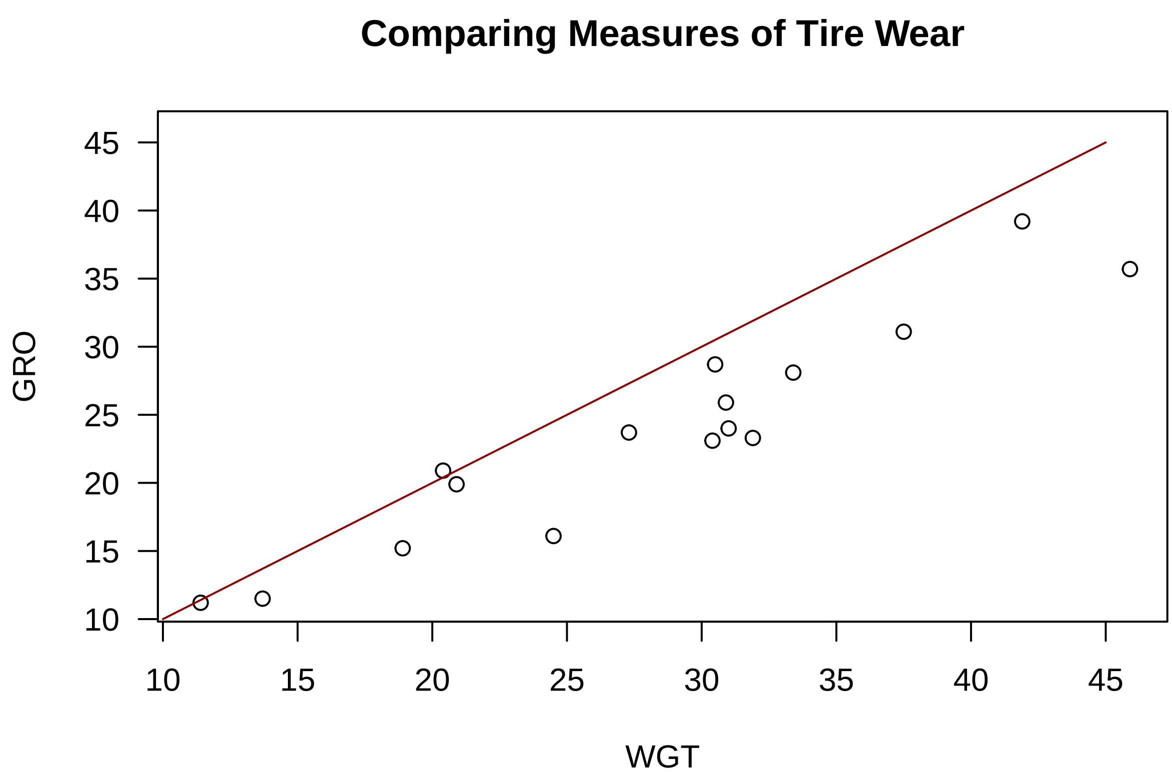
```r
# p-value = p-value = 8.504e-06 < 0.05 (reject null hypothesis, there are
# statistically significant differences between rural and non-rural salaries.)
```

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.

```r
# Read the comma-delimited text file creating a data frame object in R,
# then examine its structure:

tires <- read.csv("tires.csv")
str(tires)
```

```
## 'data.frame':    16 obs. of  2 variables:
##  $ WGT: num  45.9 41.9 37.5 33.4 31 30.5 30.9 31.9 30.4 27.3 ...
##  $ GRO: num  35.7 39.2 31.1 28.1 24 28.7 25.9 23.3 23.1 23.7 ...
```

```r
# Let's see how the measures compare on a scatter plot, using
# the same scale for both axes and a diagonal line of equality.
with(tires, plot(WGT, GRO, las = 1,
    xlim = c(min(WGT, GRO), max(WGT, GRO)),
    ylim = c(min(WGT, GRO), max(WGT, GRO))))
segments(10, 10, 45, 45, col = "darkred")  # line of equality
title("Comparing Measures of Tire Wear")
```




```r
# Note that all but one of the WGT measures is larger than
# the corresponding GRO measure.

# Now for the paired t-test
with(tires, t.test(WGT, GRO, alternative = "two.sided",
    paired = TRUE, conf.level = 0.95))
```

```
##
##  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.)
```