Class Activity#10

Sunny Lee

2021-03-10

Wilcoxon Rank Sum Tests for Two Independent Samples.

Data Set NLBB salaries

1) Using a software package, conduct the Wilcoxon rank sum test to determine if the distribution of salaries is different for pitchers than for first basemen.

```
salary_1 <- read.csv("C1 NLBB Salaries.csv")
pitcher <- salary_1[salary_1$Position == "Pitcher", ]$Salary
first_basemen <- salary_1[salary_1$Position == "First Baseman", ]$Salary
wilcox.test(pitcher, first_basemen, exact = FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  pitcher and first_basemen
## W = 3, p-value = 0.0601
## alternative hypothesis: true location shift is not equal to 0
```

```
rank(pitcher, first_basemen)
```

```
## [1] 1 2 3 4 5
```

```
test <- c(first_basemen, pitcher)
rank(test)
```

```
##  [1]  5  6  7  9 10  1  2  3  4  8
```

From the Wilcoxon rank sum test, we find that the p value we obtain is greater than .05, and thus we fail to reject the null hypothesis. Thus, we conclude that the distribution of salaries is not different for pitchers and first basemen.

2) Find $2 * P(W \leq 18)$ assuming $W \sim N(27.5, 4.787)$. How does your answer compare to that from Question 1?

```
2 * pnorm(18, mean = 27.5, sd = 4.787)
```

```
## [1] 0.04719551
```

```
#to use a table, we have to standardize it
2 * pnorm((18-27.5)/ 4.787)
```

```
## [1] 0.04719551
```

The p values obtained are less than .05, and thus we may reject the null hypothesis, which is different from what we obtained in question 1, where we failed to reject the null hypothesis.

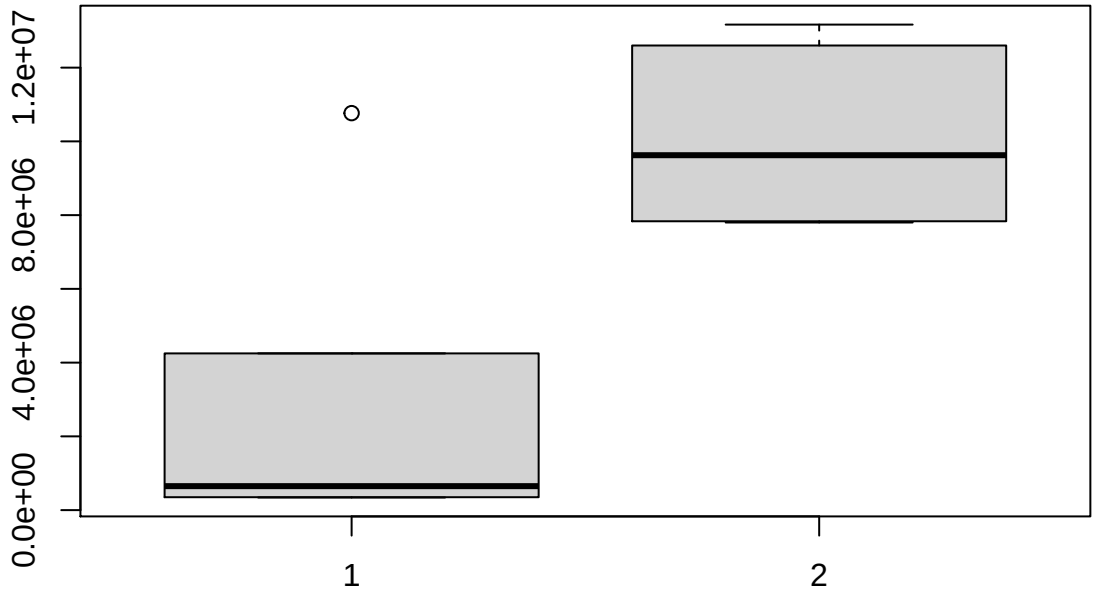
3) Use a two-sided two-sample t-test (assume unequal variance) to analyze the data. Are your conclusions the same as in question 1? Create an individual value plot of the data. Are any distributional

assumptions violated? Which test is more appropriate to use for this data set?

```
t.test(pitcher, first_basemen, equal.var = FALSE , alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  pitcher and first_basemen
## t = -2.9926, df = 6.3438, p-value = 0.02266
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12529089  -1337868
## sample estimates:
## mean of x mean of y
##   3271522  10205000
```

```
boxplot(pitcher, first_basemen)
```



Our conclusion differs from question 1, as we obtain a p value less than .05, and thus reject the null hypothesis. From the boxplot above, we see the variance and mean of the two distributions are not the same and thus we conclude the t.test is more accurate for this problem.

Kruskal-Wallis Test for Two or More Independent Samples

Data set: NLBB Salaries

4) Using a software package, run the Kruskal-Wallis test (use all three groups with samples of size 5 per group) to determine if the distribution of salaries differs by position. Create an individual value plot the data. Do the data look normally distributed in each group?

```
kruskal.test(salary_1$Salary~salary_1$Position)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  salary_1$Salary by salary_1$Position
## Kruskal-Wallis chi-squared = 7.98, df = 2, p-value = 0.0185
```

```
boxplot(salary_1$Salary~salary_1$Position)
```

From our Kruskal-Wallis test, we obtain a very low p value and reject the null hypothesis. Thus, we conclude that the three distributions are not the same, which we can see from the three boxplots above. Though catcher and pitcher seem to have a similar mean, their variances are quite different and the first basement have a drastically different mean.