# Baseball 1986 Season

## JDT

## Fall 2023

## Contents

## 1 Introduction

The Sashelp.Baseball data set contains salary and performance information for Major League Baseball players (excluding pitchers) who played at least one game in both the 1986 and 1987 seasons (Time Inc. 1987). The salaries are for the 1987 season, and the performance measures are from the 1986 season. The data set contains 322 observations.

In this document, I have used the seasonal results for the 1986 mlb season in which the National League champions, NY Mets defeated the American League champions, Boston Red Sox for the World Series Championship. The seasonal record is given below

## 1986 MLB National League East

| Team | GB |
|---|---|
| Mets | - |
| Phillies | 21.5 |
| Cardinals | 28.5 |
| Nationals | 29.5 |
| Cubs | 37.0 |
| Pirates | 44.0 |

## 1986 MLB National League West

| Team | GB |
|---|---|
| Astros | - |
| Reds | 10.0 |
| Giants | 13.0 |
| Padres | 22.0 |
| Dodgers | 23.0 |
| Braves | 23.5 |

## 1986 MLB American League East

| Team | GB |
|---|---|
| Red Sox | - |
| Yankees | 5.5 |
| Tigers | 8.5 |
| Blue Jays | 9.5 |
| Guardians | 11.5 |
| Brewers | 18.0 |
| Orioles | 22.5 |

## 1986 MLB American League West

| Team | GB |
|---|---|
| Angels | - |
| Rangers | 5.0 |
| Royals | 16.0 |
| Athletics | 16.0 |
| White Sox | 20.0 |
| Twins | 21.0 |

My objective with these data is to investigate how the RedSox compared with their inter-divisional rival NY Yankees using selected seasonal baseball statistics. In this case the number of hits for the season (nHits). This is problem 1.

The second problem (problem 2) compares the NY Mets seasonal baseball statistics with the average career statistics for each positional player on the 1986 NYMets team. Again nHits was chosen for illustrative purposes.

Once I saw the limited number of observations for these problems, I changed my comparisons. In problem 1 I compared the East versus the West division of the American league using only the 1986 data. In the second problem I compared the 1986 results with average career numbers for the National League East division.

# 2 Theory

## 2.1 Test of Hypothesis for Two Populations

This section will contain a number of procedures for testing equality of two location and scale parameters using both parametric and non parametric procedures when the two samples are independent or dependent as in the paired design assumptions. The parametric method is covered in an introductory course, such as, STAT 2381. Since, you have seen this material before I will briefly cover the material and skip parts in the interest of time and the objectives of this course.

### Independent Sample Design

The parametric method for testing hypotheses concerning the means for two normal populations is presented. Let $y_{1i} \sim N(\mu_1, \sigma_1^2)$ for $i = 1, \ldots, n_1$ denote a random sample from population 1 and $y_{2i} \sim N(\mu_2, \sigma_2^2)$ for $i = 1, \ldots, n_2$ denote a random sample from population 2 when $\sigma_1$, and $\sigma_2$ are unknown. The within group sample mean estimates ($\bar{y}_1$ and $\bar{y}_2$), sample standard deviation estimates ($s_1$ and $s_2$), standard errors ($se_1$ and $se_2$), and confidence limits for means and standard deviations are computed in the same way as for the one-sample design. The mean difference $\mu_1 - \mu_2 = \mu_d$ is estimated by $\bar{y}_d = \bar{y}_1 - \bar{y}_2$.

Under the assumption of **equal variances** ($\sigma_1^2 = \sigma_2^2$), the pooled estimate of the common standard deviation is

$$s_p = \left( \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right)^{\frac{1}{2}}$$

The pooled standard error (the estimated standard deviation of $\bar{y}_d$ assuming equal variances) is

$$se_p = s_p \left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{\frac{1}{2}}$$

The pooled $100(1 - \alpha)\%$ confidence interval for the mean difference $\mu_d$ is

$$\left( \bar{y}_d \pm t_{1-\frac{\alpha}{2}, n_1+n_2-2} \times se_p \right)$$

The $t$ value for the pooled test is computed as

$$t_p = \frac{\bar{y}_d - \mu_0}{se_p}$$

The two-sided p-value of the test is computed as

$$\Pr[t_p^2 > F_{1-\alpha, 1, n_1+n_2-2}]$$

3

Under the assumption of **unequal variances** (called the the Behrens-Fisher problem), the un-pooled standard error is computed as

$$se_u = \left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^{\frac{1}{2}}$$

Satterthwaite's (1946) approximation for the degrees of freedom is

$$df_u = \frac{se_u^4}{\frac{s_1^4}{(n_1-1)n_1^2} + \frac{s_2^4}{(n_2-1)n_2^2}}$$

The unpooled Satterthwaite $100(1-\alpha)\%$ confidence interval for the mean difference $\mu_d$ is

$$\left( \bar{y}_d \pm t_{1-\frac{\alpha}{2},df_u} se_u \right)$$

The $t$ value for the unpooled Satterthwaite test is computed as

$$t_u = \frac{\bar{y}_d - \mu_0}{se_u}$$

The two-sided p-value of the unpooled Satterthwaite test is computed as

$$\Pr[t_u > F_{1-\alpha,1,df_u}]$$

When the COCHRAN option is specified in the PROC TTEST statement, the Cochran and Cox (1950) approximation of the p-value of the $t_u$ statistic is the value of p such that

$$t_u = \frac{\left( \frac{s_1^2}{n_1} \right) t_1 + \left( \frac{s_2^2}{n_2} \right) t_2}{\left( \frac{s_1^2}{n_1} \right) + \left( \frac{s_2^2}{n_2} \right)}$$

where $t_1$ and $t_2$ are the critical values of the $t$ distribution corresponding to a significance level of p and sample sizes of $n_1$ and $n_2$, respectively. The number of degrees of freedom is undefined when $n_1 \neq n_2$. In general, the Cochran and Cox test tends to be conservative (Lee and Gurland, 1975).

The $100(1-\alpha)\%$ CI=EQUAL and CI=UMPU confidence intervals for the common population standard deviation $\sigma$ assuming equal variances are computed as discussed in the section Normal Data (DIST=NORMAL) for the one-sample design, except replacing $s^2$ by $s_p^2$ and $(n-1)$ by $(n_1 + n_2 - 1)$.

The folded form of the F statistic, $F'$, tests the hypothesis that the variances are equal (Steel and Torrie, 1980), where

$$F' = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)}$$

The p-value gives the probability of a greater F value under the null hypothesis that $\sigma_1^2 = \sigma_2^2$.[1]

## 2.2 Paired Tests

The analysis is the same as the analysis for the one-sample design in the section Normal Data (DIST=NORMAL) based on the differences

$$d_i = y_{1i} - y_{2i}, \quad i \in \{1, \ldots, n\}.$$

---

[1]This test is not very robust to violations of the assumption that the data are normally distributed, and thus it is not recommended without confidence in the normality assumption.

## 2.3 Simple Linear Rank Tests for Two-Sample Data

The material in this section may be new for you. It provides the needed theory for many of the non parametric methods for comparing parameters, such the median, from two populations that are not normally distributed.

Statistics of the form

$$S = \sum_{j=1}^{n} a(R_j)$$

are called *simple linear rank statistics*, where $R_j$ is the rank of observation $j$, $a(R_j)$ is the score based on the rank of observation $j$, and $n$ is the total number of observations[2].

To compute an asymptotic test for a linear rank sum statistic, use the standardized test statistic $z$, which has an asymptotic standard normal distribution under the null hypothesis as

$$z = \frac{(S - E_0(S))}{\sqrt{Var_0(S)}}$$

where $E_0(S)$ is the expected value of $S$ under the null hypothesis, and $Var_0(S)$ is the variance under the null hypothesis. As shown in Randles and Wolfe (1979),

$$E_0(S) = \frac{n_1}{n} \sum_{j=1}^{n} a(R_j)$$

where $n_1$ is the number of observations in the first class level (sample), $n_2$ is the number of observations in the other class level, and

$$Var_0(S) = \frac{n_1 n_2}{n(n-1)} \sum_{j=1}^{n} (a(R_j) - \bar{a})^2$$

where $\bar{a}$ is the average score,

$$\bar{a} = \frac{1}{n} \sum_{j=1}^{n} a(R_j)$$

## 2.4 Scores for Linear Rank Tests

The following score types are used primarily to test for differences in:

- Location
  - Wilcoxon, median, Van der Waerden (normal), Savage, and Conover.
- Scale
  - Siegel-Tukey, Ansari-Bradley, Klotz, Mood, and Conover.

Conover scores can be used to test for differences in both location and scale. This section gives formulas for the score types. For further information about the formulas and the applicability of each score, see Randles and Wolfe (1979), Gibbons and Chakraborti (2010), Conover (1999), and Hollander and Wolfe (1999).

---

[2]For two-sample data (where the observations are classified into two levels), PROC NPAR1WAY calculates simple linear rank statistics for the scores that you specify.

## Wilcoxon Scores

Wilcoxon scores are the ranks of the observations, $a(R_j) = R_j$ where $R_j$ is the rank of observation $j$. The Wilcoxon scores in the linear rank statistic for two-sample data are used to perform the rank sum statistic for the Mann-Whitney-Wilcoxon test. The Wilcoxon scores are used in the one-way ANOVA statistic to perform the Kruskal-Wallis test.

## Median Scores

Median scores equal 1 for observations greater than the median, and 0 otherwise. In terms of the observation ranks, median scores are defined as

$$a(R_j) = \begin{cases} 1 & \text{if} \quad R_j > (n+1)/2 \\ 0 & \text{if} \quad R_j \leq (n+1)/2 \end{cases}$$

Use the median scores in the linear rank statistic for two-sample data to produce the two-sample median test. The one-way ANOVA statistic with median scores is equivalent to the Brown-Mood test. Median scores are particularly powerful for distributions that are symmetric and heavy-tailed.

## Van der Waerden (Normal) Scores

Van der Waerden scores are the quantiles of a standard normal distribution and are also known as *quantile normal* scores. Van der Waerden scores are computed as

$$a(R_j) = \Phi^{-1}\left(\frac{R_j}{n+1}\right)$$

where $\Phi$ is the cumulative distribution function of a standard normal distribution. These scores are powerful for normal distributions.

## Savage Scores

Savage scores are expected values of order statistics from the exponential distribution, with 1 subtracted to center the scores around 0. Savage scores are computed as

$$a(R_j) = \sum_{i=1}^{R_j}\left(\frac{1}{n-i+1}\right) - 1$$

Savage scores are powerful for comparing scale differences in exponential distributions or location shifts in extreme value distributions (Hajek, 1969, p. 83).

## Siegel-Tukey Scores

$$\begin{aligned} a(1) &= 1, \quad a(n) = 2, \quad &a(n-1) = 3, \quad a(2) = 4, \\ a(3) &= 5, \quad a(n-2) = 6, \quad &a(n-3) = 7, \quad a(4) = 8, \ \ldots \end{aligned}$$

where the score values continue to increase in this pattern toward the middle ranks until all observations have been assigned a score.

## Ansari-Bradley Scores

Ansari-Bradley scores are similar to Siegel-Tukey scores, but Ansari-Bradley scoring assigns the same score value to corresponding extreme ranks. The Siegel-Tukey scores are a permutation of the ranks $1, 2, \ldots, n$. Ansari-Bradley scores are defined as

$$a(1) = 1, \quad a(n) = 1,$$
$$a(2) = 2, \quad a(n-1) = 2, \ \ldots$$

Equivalently, Ansari-Bradley scores are equal to

$$a(R_j) = \frac{n+1}{2} - \left| R_j - \frac{n+1}{2} \right|$$

## Klotz Scores

Klotz scores are the squares of the Van der Waerden (normal) scores. Klotz scores are computed as

$$a(R_j) = \left( \Phi^{-1} \left( \frac{R_j}{n+1} \right) \right)^2$$

where $\Phi$ is the cumulative distribution function of a standard normal distribution.

## Mood Scores

Mood scores are computed as the square of the difference between the observation rank and the average rank. Mood scores can be written as

$$a(R_j) = \left( R_j - \frac{n+1}{2} \right)^2$$

## Conover Scores

Conover scores are based on the squared ranks of the absolute deviations from the sample means. For observation $j$ the absolute deviation from the mean is computed as

$$U_j = |X_{j(i)} - \bar{X}_i|$$

where $X_{j(i)}$ is the value of observation $j$, observation $j$ belongs to sample $i$, and $\bar{X}_i$ is the mean of sample $i$. The values of $U_j$ are ranked, and the Conover score for observation $j$ is computed as

$$\text{Score}_j = (\text{Rank}(U_j))^2$$

The Conover score test is also known as the squared ranks test for variances. See Conover (1999) for more information.

## 2.5 Wilcoxon and Mann-Whitney Test

The Mann-Whitney and Wilcoxon test assumes that

- The data consists of a random sample of $n_1$ values, denoted $X_1, X_2, \ldots, X_{n_1}$ from $F_X(x)$ with $median(X) = \theta_X$, and a random sample of $n_2$ values, denoted $Y_1, Y_2, \ldots, Y_{n_2}$ from $F_Y(y)$ with $median(Y) = \theta_Y$.

- The random samples are at least ordinal.

- $F_X$ and $F_Y$ differ only with respect to the their median. That is, $\theta_X = \theta_Y + \delta$.

The null hypothesis is $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ (the usual one-sided tests are possible). The procedure is

1. Combine the data for the two random samples and rank the combined data where $r_j = rank(Y_j)$.

2. Compute the Wilcoxon statistics as, $W = \sum_{j=1}^{n_2} r_j$.

3. Reject $H_0$ if $W$ is either too small ($\theta_Y < \theta_X$) or too large ($\theta_Y > \theta_X$).

4. The large sample distribution of $W^* = \frac{W - E(W)}{\sqrt{Var(W)}} \sim N(0, 1)$, where

$$E(W) = \frac{n_2(N + 1)}{2}$$

and

$$Var(W) = \frac{n_1 n_2(N + 1)}{12}$$

for $N = n_1 + n_2$.

The Mann-Whitney test statistics is

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi(X_i, Y_j)$$

where $\phi(x, y) = I_{x<y}$ the indicator function. The two statistics are similar in that

$$W = U + \frac{n_2(n_2 + 1)}{2}.$$

Note: $\Pr[X < Y] = E(I_{X<Y})$ in which case the statistic $U$ can be used as a nonparametric estimate of $\Pr[X < Y]$.[3]

---

[3]Since $\Pr[X < Y]$ = AUC, the Mann-Whitney statistics is a nonparametric estimate for the area under the ROC curve (AUC). Delong presented a method for computing the standard error for the Mann-whitney statistic.

## 2.6 Tests Based on the Empirical Distribution Function (EDF)

This section describes three nonparametric tests that are based on the empirical distribution function[4]. The procedures are; the Kolmogorov-Smirnov and Cramer-von Mises tests, and also the Kuiper test for two-sample data.[5] The null hypothesis is $H_0 : F_X(\cdot) = F_Y(\cdot) = F(\cdot)$. The (EDF) of a sample $\{x_j\}$, $j = 1, 2, \ldots, n$ is defined as

$$\hat{F}(x) = \frac{1}{n}(\text{number of } x_j \leq x) = \frac{1}{n}\sum_{j=1}^{n} I(x_j \leq x)$$

where $I(\cdot)$ is an indicator function. Let $\hat{F}_i$ denote the sample EDF for the $i^{th}$ group. The EDF for the overall sample, pooled over groups, can also be expressed as

$$\hat{F}(x) = \frac{1}{n}\sum_{i} \left(n_i \hat{F}_i(x)\right)$$

where $n_i$ is the number of observations in the $i^{th}$ group, and $n$ is the total number of observations.

## Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov statistic measures the maximum deviation of the EDF within the groups from the pooled EDF. The Kolmogorov-Smirnov statistic is computed as,

$$KS = \max_{j} \sqrt{\frac{1}{n}\sum_{i} n_i \left(\hat{F}_i(x_j) - \hat{F}(x_j)\right)^2} \quad \text{for} \quad j = 1, 2, \ldots, n$$

The asymptotic Kolmogorov-Smirnov statistic is computed as, $KS_a = KS \times \sqrt{n}$. If there are only two class levels, the two-sample Kolmogorov-Smirnov test statistic D is

$$D = \max_{j} \left|\hat{F}_1(x_j) - \hat{F}_2(x_j)\right| \quad \text{for} \quad j = 1, 2, \ldots, n$$

The p-value for this test is the probability that D is greater than the observed value d under the null hypothesis of no difference between class levels (samples). The asymptotic p-value for D is approximated as,

$$\Pr(D > d) = 2 \sum_{i=1}^{\infty}(-1)^{(i-1)} e^{(-2i^2 z^2)}$$

where

$$z = d\sqrt{n_1 n_2/n}$$

See Hodges (1957) for information about this approximation.

---

[4][PROC NPAR1WAY - EDF option].

[5]For further information about the formulas and the interpretation of EDF statistics, see Hollander and Wolfe (1999) and Gibbons and Chakraborti (2010). For details about the $k$-sample analogs of the Kolmogorov-Smirnov and Cramer-von Mises statistics, see Kiefer (1959).

### Cramer-von Mises Test

The Cramer-von Mises statistic is

$$CM = \frac{1}{n^2} \sum_i \left( n_i \sum_{j=1}^{p} t_j \left( \hat{F}_i(x_j) - \hat{F}(x_j) \right)^2 \right)$$

where $t_j$ is the number of ties at the $j^{th}$ distinct value and p is the number of distinct values. The asymptotic value is computed as

$$CM_a = CM \times n.$$

### Kuiper Test

For data with two class levels, the Kuiper statistic is

$$K = \max_j \left( \hat{F}_1(x_j) - \hat{F}_2(x_j) \right) - \min_j \left( \hat{F}_1(x_j) - \hat{F}_2(x_j) \right) \quad \text{where} \quad j = 1, 2, \ldots, n$$

The asymptotic value is

$$K_a = K \sqrt{n_1 n_2 / n}$$

The p-value for the Kuiper test is the probability of observing a larger value of $K_a$ under the null hypothesis of no difference between the two classes Owen (1962, p 441).

# 3   SAS

## 3.1   Header Code

```
options center nodate pagesize=80 ls=70;
libname ldata '/home/jacktubbs/my_shared_file_links/jacktubbs/LaTeX/Class';

/* Simplified LaTeX output that uses plain LaTeX tables  *
ods latex path='/home/jacktubbs/my_shared_file_links/jacktubbs/LaTeX/clean'
 file='baseball_4382.tex' style=journal
stylesheet="sas.sty"(url="sas");
*/

/*
http://support.sas.com/rnd/base/ods/odsmarkup/latex.html
*
ods graphics / reset width=5in  outputfmt=png
  antialias=on;
*/;

title "1986 Baseball Data";
/*
In 1986 mlb consisted of two leagues - American and National -
each with two divisions - East and West. There was not a wildcard
```

```
so each team played 162 games to determine the 4 league/divisional winners
which each played a best of 7 to determine the league champion who
then played a best of 7 to determine the world champion!  The American league
used the designated hitter and the National league did not.

The divisional leaders were NY Mets and Houston Astros in the National league.
The divisional leaders were Boston RSox and LA Angels in the American league.

In this analysis I have two problem in mind.
        1. Compare teams from the same division using 1986 data.
        2. Compare the performance of a team versus their career stats (yearly average
           over each players career)


*/

data baseball; set sashelp.baseball;
run;


data baseball; set baseball;
in_fielder = (position in ('1B' '2B' 'SS' '3B'));
out_fielder = (position in ('CF' 'RF' 'LF' 'OF'));
catcher = (position = 'C');
CrHits2 = CrHits*CrHits;
run;

proc freq data=baseball; table league*division; run;
```

## 3.2   Output

### 3.2.1   Problem 1

```
/********************************;
Problem 1 – comparing the American League East versus
the West using 1986 seasonal data
*******************************/;

title3 'Comparing Divisions in the American League';
data problem1; set baseball;
if league = 'American';
*if team in ('New York', 'Boston');
keep  Salary Team Division YrMajor logSalary nAtBat
nBB nError nHits nHome nRBI nRuns;
run;

proc freq data=problem1; table division; run;
```

```
proc sort data=problem1; by division; run;

title3 'Test for means assuming normal data';
title4 'Using Number of Hits';

proc ttest data=problem1 alpha=.05; class division;
var nHits;                         /*   Use which variable you choose  */;
run;

title3 'Test for means assuming non-normal data';
proc npar1way data=problem1 wilcoxon edf normal; class division;
var nHits;                         /*   Use which variable you choose  */;
run;
```

### 1986 Baseball Data

### The FREQ Procedure

| League | Table of League by Division | | |
|---|---|---|---|
| | Division | | |
| | East | West | Total |
| American | 85<br>26.40<br>48.57<br>54.14 | 90<br>27.95<br>51.43<br>54.55 | 175<br>54.35 |
| National | 72<br>22.36<br>48.98<br>45.86 | 75<br>23.29<br>51.02<br>45.45 | 147<br>45.65 |
| Total | 157<br>48.76 | 165<br>51.24 | 322<br>100.00 |

### The FREQ Procedure

| Division at the End of 1986 | | | | |
|---|---|---|---|---|
| Division | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| East | 85 | 48.57 | 85 | 48.57 |
| West | 90 | 51.43 | 175 | 100.00 |

***Test for means assuming normal data***

***Using Number of Hits***

***The TTEST Procedure***

***Variable: nHits (Hits in 1986)***

| Division | Method | N | Mean | Std Dev | Std Err | Minimum | Maximum |
|----------|--------|---|------|---------|---------|---------|---------|
| East | | 85 | 113.3 | 48.7724 | 5.2901 | 36.0000 | 238.0 |
| West | | 90 | 102.4 | 41.1687 | 4.3396 | 39.0000 | 223.0 |
| Diff (1-2) | Pooled | | 10.9739 | 45.0213 | 6.8094 | | |
| Diff (1-2) | Satterthwaite | | 10.9739 | | 6.8423 | | |

| Division | Method | Mean | 95% CL Mean | | Std Dev | 95% CL Std Dev | |
|----------|--------|------|-------------|------|---------|----------------|------|
| East | | 113.3 | 102.8 | 123.8 | 48.7724 | 42.3818 | 57.4504 |
| West | | 102.4 | 93.7329 | 111.0 | 41.1687 | 35.9078 | 48.2499 |
| Diff (1-2) | Pooled | 10.9739 | −2.4663 | 24.4140 | 45.0213 | 40.7360 | 50.3223 |
| Diff (1-2) | Satterthwaite | 10.9739 | −2.5361 | 24.4838 | | | |

| Method | Variances | DF | t Value | Pr > \|t\| |
|--------|-----------|----|---------|-----------|
| Pooled | Equal | 173 | 1.61 | 0.1089 |
| Satterthwaite | Unequal | 164.7 | 1.60 | 0.1107 |

| Equality of Variances | | | | |
|--------|--------|--------|---------|--------|
| Method | Num DF | Den DF | F Value | Pr > F |
| Folded F | 84 | 89 | 1.40 | 0.1159 |

Distribution of nHits

Q-Q Plots of nHits

**1986 Baseball Data**

**Test for means assuming non-normal data**

**The NPAR1WAY Procedure**

| Wilcoxon Scores (Rank Sums) for Variable nHits | | | Classified by Variable Division | | |
|---|---|---|---|---|---|
| Division | N | Sum of Scores | Expected Under H0 | Std Dev Under H0 | Mean Score |
| East | 85 | 7943.50 | 7480.0 | 334.936620 | 93.452941 |
| West | 90 | 7456.50 | 7920.0 | 334.936620 | 82.850000 |
| Average scores were used for ties. | | | | | |

| Wilcoxon Two-Sample Test | | | | | |
|---|---|---|---|---|---|
| Statistic | Z | Pr > Z | Pr > \|Z\| | t Approximation | |
| | | | | Pr > Z | Pr > \|Z\| |
| 7943.500 | 1.3824 | 0.0834 | 0.1669 | 0.0843 | 0.1686 |
| Z includes a continuity correction of 0.5. | | | | | |

| Kruskal-Wallis Test | | |
|---|---|---|
| Chi-Square | DF | Pr > ChiSq |
| 1.9150 | 1 | 0.1664 |
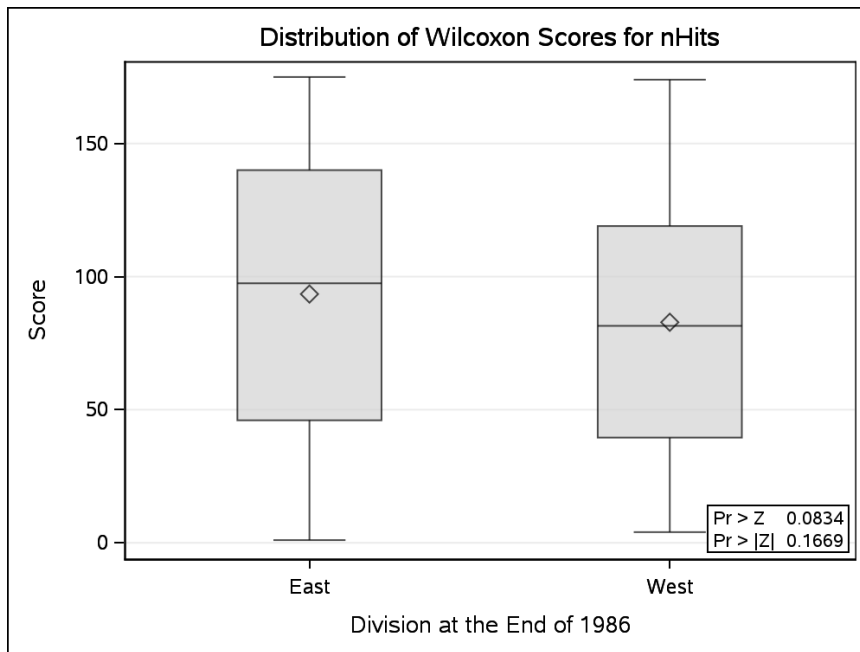
15

Distribution of Wilcoxon Scores for nHits

| | |
|---|---|
| Pr > Z | 0.0834 |
| Pr > \|Z\| | 0.1669 |

**Test for means assuming non-normal data**

**The NPAR1WAY Procedure**

| Van der Waerden Scores (Normal) for Variable nHits | | | | Classified by Variable Division | |
|---|---|---|---|---|---|
| Division | N | Sum of Scores | Expected Under H0 | Std Dev Under H0 | Mean Score |
| East | 85 | 9.046291 | 0.0 | 6.461690 | 0.106427 |
| West | 90 | −9.046291 | 0.0 | 6.461690 | −0.100514 |
| Average scores were used for ties. | | | | | |

| Van der Waerden Two-Sample Test | | | |
|---|---|---|---|
| Statistic | Z | Pr > Z | Pr > \|Z\| |
| 9.0463 | 1.4000 | 0.0808 | 0.1615 |

| Van der Waerden One-Way Analysis | | |
|---|---|---|
| Chi-Square | DF | Pr > ChiSq |
| 1.9600 | 1 | 0.1615 |
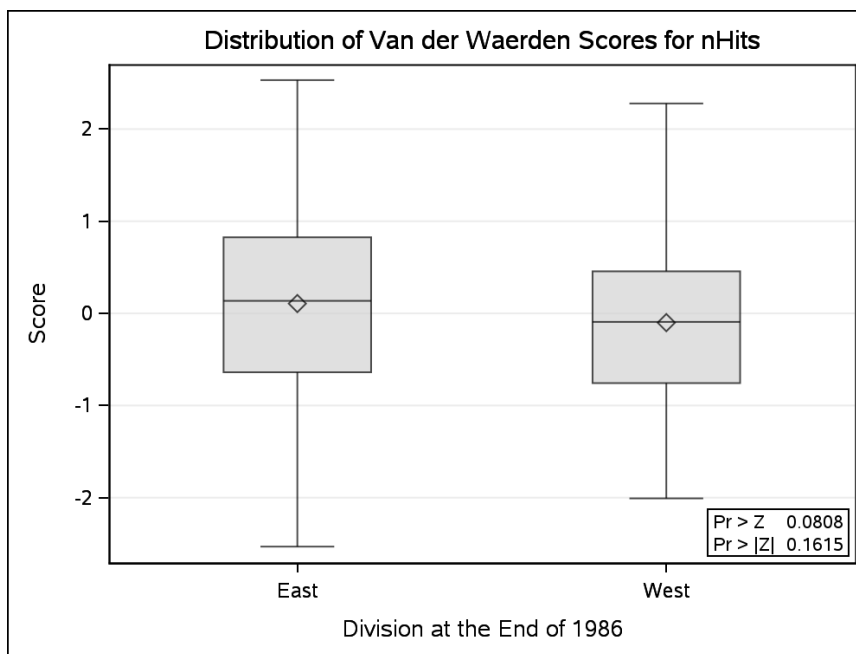


Distribution of Van der Waerden Scores for nHits

17

**Test for means assuming non-normal data**

**The NPAR1WAY Procedure**

| Kolmogorov-Smirnov Test for Variable nHits Classified by Variable Division | | | |
|---|---|---|---|
| Division | N | EDF at Maximum | Deviation from Mean at Maximum |
| East | 85 | 0.611765 | −0.681781 |
| West | 90 | 0.755556 | 0.662572 |
| Total | 175 | 0.685714 | |
| Maximum Deviation Occurred at Observation 102 | | | |
| Value of nHits at Maximum = 131.0 | | | |

| Kolmogorov-Smirnov Two-Sample Test (Asymptotic) | | | |
|---|---|---|---|
| KS | 0.071866 | D | 0.143791 |
| KSa | 0.950699 | Pr > KSa | 0.3266 |



Empirical Distribution for nHits
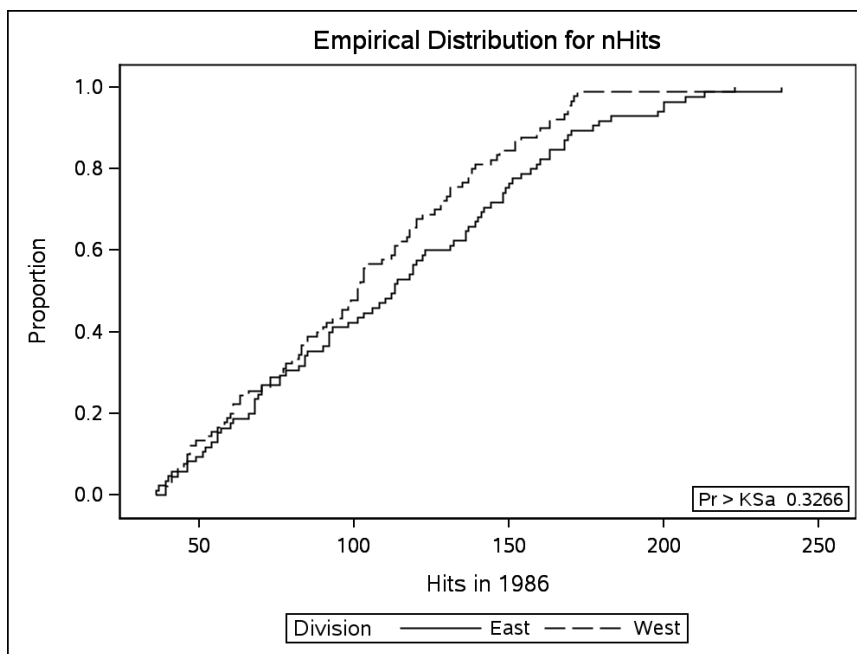
| Cramer-von Mises Test for Variable nHits Classified by Variable Division | | | |
|---|---|---|---|
| Division | N | Summed Deviation | from Mean |
| East | 85 | | 0.126974 |
| West | 90 | | 0.119920 |

| Cramer-von Mises Statistics (Asymptotic) | | | |
|---|---|---|---|
| CM | 0.001411 | CMa | 0.246895 |

| Kuiper Test for Variable nHits Classified by Variable Division | | | |
|---|---|---|---|
| Division | N | Deviation | from Mean |
| East | 85 | | 0.024837 |
| West | 90 | | 0.143791 |

| Kuiper Two-Sample Test (Asymptotic) | | | | | |
|---|---|---|---|---|---|
| K | 0.168627 | Ka | 1.114910 | Pr > Ka | 0.6631 |

### 3.2.2 Summary for Problem 1

When comparing the number of hits for American League Divisions there was little reason to believe that there was a difference in the means and medians for each division. The normal distribution assumption was probably okay.

### 3.2.3 Problem 2

```
/********************************;
Problem 2 - comparing the 1986 the National League East with
historical NL East teams using career averages for each
variables
********************************/;

title3 'Comparing the National League East with Career numbers';
data temp1; set baseball;
if div = 'NE';
career = 'No ';
keep  CrAtBat CrBB CrHits CrHits2 CrHome CrRbi CrRuns Div
Salary YrMajor    nBB   nHits nHome
nRBI nRuns career;
run;

data temp2; set temp1;
career = 'Yes';
nbb=CRbb/yrmajor;
nhits = crhits/yrmajor;
nhome = CRhome/yrmajor;
nrbi = CRrbi/yrmajor;
nruns = CRruns/yrmajor;
keep  CrAtBat CrBB CrHits CrHits2 CrHome CrRbi CrRuns Div
Salary YrMajor    nBB   nHits nHome
nRBI nRuns career;
run;

data problem2; set temp1 temp2; run;
proc sort data=problem2; by career; run;
proc freq data=problem2; table career; run;

title3 'Test for means assuming normal data';
proc ttest data=problem2 alpha=.05; class career;
var nHits;                        /*   Use which variable you choose  */;
run;

title3 'Test for means assuming non-normal data';
proc npar1way data=problem2 wilcoxon edf normal; class career;
var nHits;                        /*   Use which variable you choose  */;
run;
 quit;
```

*1986 Baseball Data*


*Comparing the National League East with Career numbers*

*The FREQ Procedure*

20

| career | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|--------|-----------|---------|----------------------|--------------------|
| No | 72 | 50.00 | 72 | 50.00 |
| Yes | 72 | 50.00 | 144 | 100.00 |

**Test for means assuming normal data**

**The TTEST Procedure**

**Variable: nHits (Hits in 1986)**

| career | Method | N | Mean | Std Dev | Std Err | Minimum | Maximum |
|--------|--------|---|------|---------|---------|---------|---------|
| No | | 72 | 99.8056 | 43.8055 | 5.1625 | 31.0000 | 194.0 |
| Yes | | 72 | 84.3855 | 39.2831 | 4.6296 | 16.4000 | 161.2 |
| Diff (1-2) | Pooled | | 15.4200 | 41.6058 | 6.9343 | | |
| Diff (1-2) | Satterthwaite | | 15.4200 | | 6.9343 | | |

| career | Method | Mean | 95% CL Mean | | Std Dev | 95% CL Std Dev | |
|--------|--------|------|-------------|---|---------|----------------|---|
| No | | 99.8056 | 89.5118 | 110.1 | 43.8055 | 37.6353 | 52.4145 |
| Yes | | 84.3855 | 75.1545 | 93.6166 | 39.2831 | 33.7499 | 47.0033 |
| Diff (1-2) | Pooled | 15.4200 | 1.7122 | 29.1278 | 41.6058 | 37.2783 | 47.0788 |
| Diff (1-2) | Satterthwaite | 15.4200 | 1.7108 | 29.1292 | | | |

| Method | Variances | DF | t Value | Pr > \|t\| |
|--------|-----------|-----|---------|-----------|
| Pooled | Equal | 142 | 2.22 | 0.0277 |
| Satterthwaite | Unequal | 140.35 | 2.22 | 0.0278 |

| Equality of Variances | | | | |
|-----------------------|---------|--------|---------|--------|
| Method | Num DF | Den DF | F Value | Pr > F |
| Folded F | 71 | 71 | 1.24 | 0.3607 |
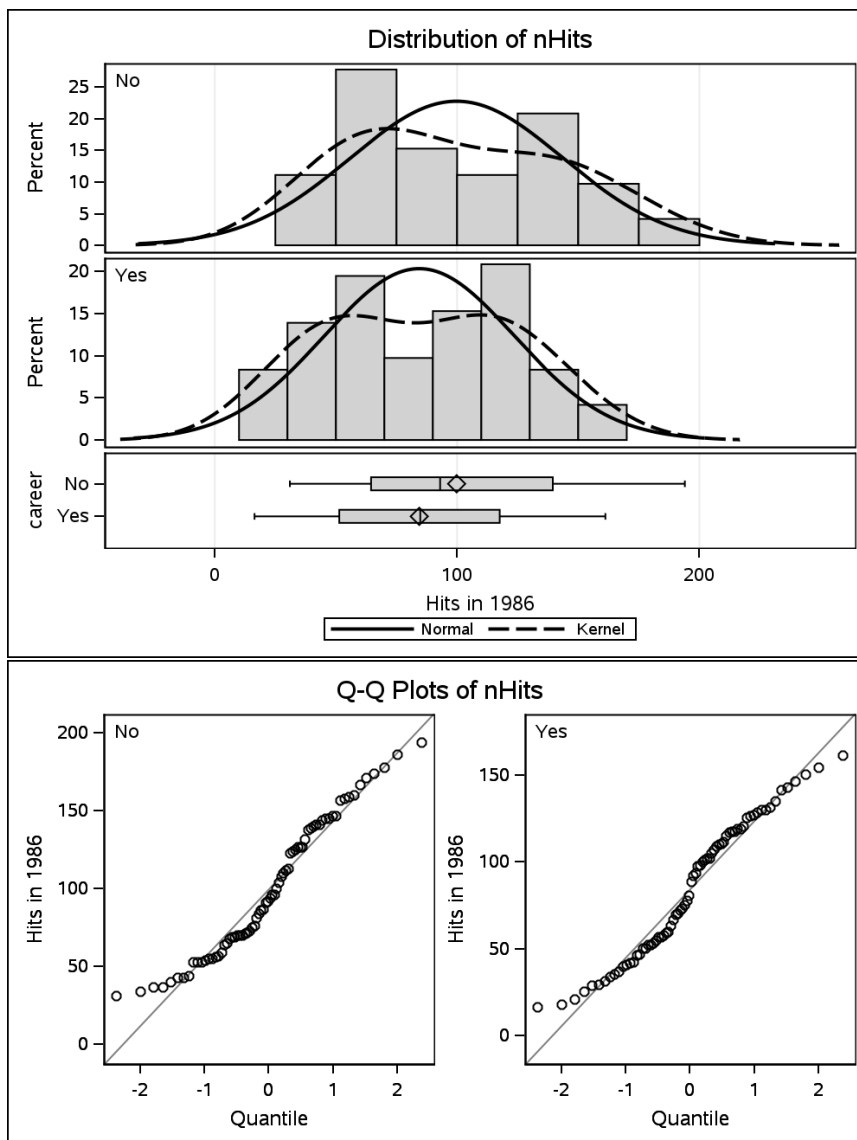
Distribution of nHits

Q-Q Plots of nHits
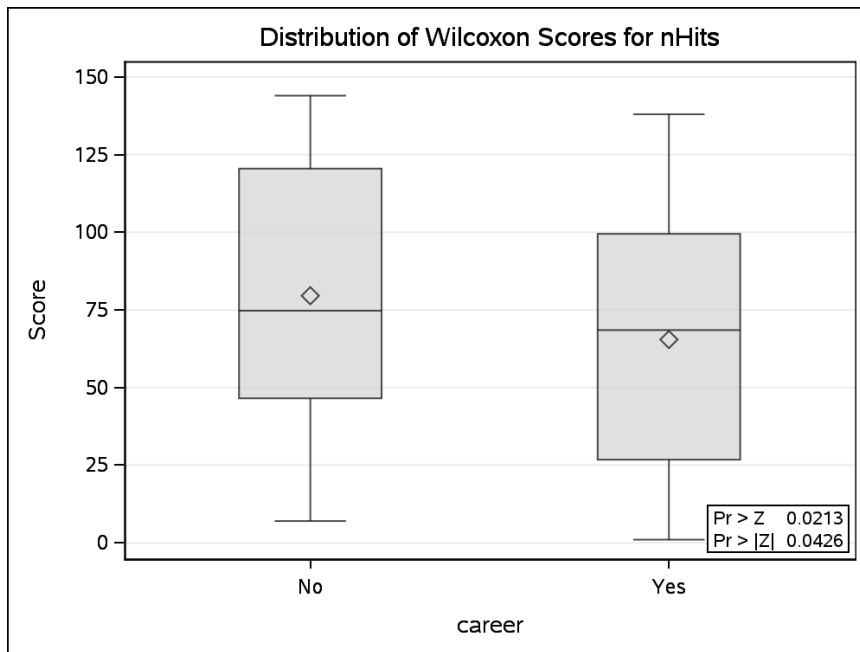
**1986 Baseball Data**

**Test for means assuming non-normal data**

**The NPAR1WAY Procedure**

| Wilcoxon Scores (Rank Sums) for Variable nHits | | | Classified by Variable career | | |
|---|---|---|---|---|---|
| career | N | Sum of Scores | Expected Under H0 | Std Dev Under H0 | Mean Score |
| No | 72 | 5728.0 | 5220.0 | 250.272048 | 79.555556 |
| Yes | 72 | 4712.0 | 5220.0 | 250.272048 | 65.444444 |
| Average scores were used for ties. | | | | | |

| Wilcoxon Two-Sample Test | | | | | |
|---|---|---|---|---|---|
| Statistic | Z | Pr > Z | Pr > |Z| | t Approximation | |
| | | | | Pr > Z | Pr > |Z| |
| 5728.000 | 2.0278 | 0.0213 | 0.0426 | 0.0222 | 0.0444 |
| Z includes a continuity correction of 0.5. | | | | | |

| Kruskal-Wallis Test | | |
|---|---|---|
| Chi-Square | DF | Pr > ChiSq |
| 4.1201 | 1 | 0.0424 |

Distribution of Wilcoxon Scores for nHits

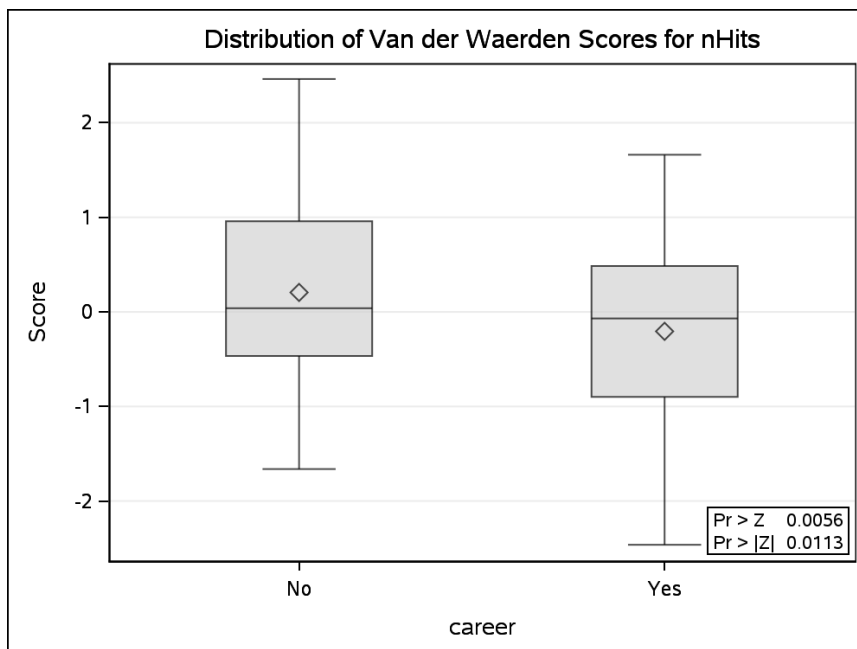| Pr > Z | 0.0213 |
| Pr > \|Z\| | 0.0426 |

*1986 Baseball Data*

*Test for means assuming non-normal data*

*The NPAR1WAY Procedure*

| Van der Waerden Scores (Normal) for Variable nHits | | | | Classified by Variable career | |
|---|---|---|---|---|---|
| *career* | *N* | *Sum of Scores* | *Expected Under H0* | *Std Dev Under H0* | *Mean Score* |
| *No* | 72 | 14.805985 | 0.0 | 5.842762 | 0.205639 |
| *Yes* | 72 | −14.805985 | 0.0 | 5.842762 | −0.205639 |
| *Average scores were used for ties.* | | | | | |

| Van der Waerden Two-Sample Test | | | |
|---|---|---|---|
| *Statistic* | *Z* | *Pr > Z* | *Pr > |Z|* |
| 14.8060 | 2.5341 | 0.0056 | 0.0113 |

| Van der Waerden One-Way Analysis | | |
|---|---|---|
| *Chi-Square* | *DF* | *Pr > ChiSq* |
| 6.4215 | 1 | 0.0113 |



Distribution of Van der Waerden Scores for nHits

25

**Test for means assuming non-normal data**

**The NPAR1WAY Procedure**

| Kolmogorov-Smirnov Test for Variable nHits Classified by Variable career | | | |
|---|---|---|---|
| career | N | EDF at Maximum | Deviation from Mean at Maximum |
| No | 72 | 0.708333 | −0.824958 |
| Yes | 72 | 0.902778 | 0.824958 |
| Total | 144 | 0.805556 | |
| Maximum Deviation Occurred at Observation 89 | | | |
| Value of nHits at Maximum = 131.466667 | | | |

| Kolmogorov-Smirnov Two-Sample Test (Asymptotic) | | | |
|---|---|---|---|
| KS | 0.097222 | D | 0.194444 |
| KSa | 1.166667 | Pr > KSa | 0.1314 |



Empirical Distribution for nHits
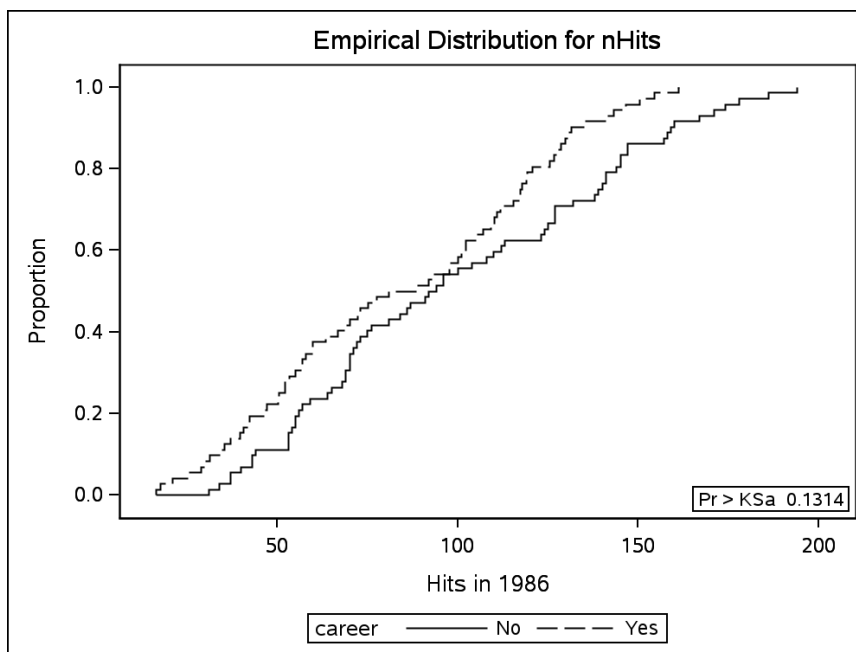
| Cramer-von Mises Test for Variable nHits Classified by Variable career | | | |
|---|---|---|---|
| career | N | Summed Deviation | from Mean |
| No | 72 | | 0.203125 |
| Yes | 72 | | 0.203125 |

| Cramer-von Mises Statistics (Asymptotic) | | | |
|---|---|---|---|
| CM | 0.002821 | CMa | 0.406250 |

| Kuiper Test for Variable nHits Classified by Variable career | | | |
|---|---|---|---|
| career | N | Deviation | from Mean |
| No | 72 | | 0.000000 |
| Yes | 72 | | 0.194444 |

| Kuiper Two-Sample Test (Asymptotic) | | | | | |
|---|---|---|---|---|---|
| K | 0.194444 | Ka | 1.166667 | Pr > Ka | 0.5850 |

### 3.2.4  Summary for Problem 2

When comparing the number of hits for the 1986 NL East team with their average career numbers reveal a difference in the means and medians. The normal distribution assumption was questionable in that both density were bimodal with differing centers. I did not check for goodness of fit but the EDF approach reveal difference in the two populations. This is expected since the NL East did not win the world series every year!