NASCAR Winston Cup Race Results for 1975-2003

Larry Winner
University of Florida

*Journal of Statistics Education* Volume 14, Number 3 (2006),
ww2.amstat.org/publications/jse/v14n3/datasets.winner.html

Copyright © 2006 by Larry Winner, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

**Key Words:** Kendall’s τ; Matched pairs; Ordinal data; Spearman’s ρ; Sports statistics.

**Abstract**

Stock car racing has seen tremendous growth in popularity in recent years. We introduce two datasets containing results from all Winston Cup races between 1975 and 2003, inclusive. Students can use any number of statistical methods and applications of basic probability on the data to answer a wide range of practical questions. Instructors and students can define many types of events and obtain their corresponding empirical probabilities, as well as gain a hands-on computer-based understanding of conditional probabilities and probability distributions. They can model the rapid growth of the sport based on total payouts by year in real and adjusted dollars, applying linear and exponential growth models that are being taught at earlier stages in introductory statistics courses. Methods of making head-to-head comparisons among pairs of drivers are demonstrated based on their start and finish order, applying a simple to apply categorical method based on matched pairs that students can easily understand, but may not be exposed to in traditional introductory methods courses. Spearman’s and Kendall’s rank correlation measures are applied to each race to describe the association between starting and finishing positions among drivers, which students can clearly understand are ordinal, as opposed to interval scale outcomes. A wide variety of other potential analyses may also be conducted and are briefly described. The dataset nascard.dat.txt is at the driver/race level and contains variables including: driver name, start and finish positions, car make, laps completed, and prize winnings. The dataset nascarr.dat.txt is at the race level and contains variables including: number of drivers, total prize money, monthly consumer price index, track length, laps completed, numbers of caution flags and lead changes, completion time, and spatial coordinates of the track. These datasets offer students and instructors many opportunities to explore diverse statistical applications.

**1. Introduction**

The National Association for Stock Car Auto Racing (NASCAR) was initiated in December, 1947 when Bill France met with regional promoters in Daytona Beach, Florida to lay out rules and regulations for the profession (all NASCAR history is obtained from *NASCAR Record & Fact Book (2004 Ed.)*). The first race was held February 15, 1948 in Daytona Beach, FL. In June, 1949 the first race among automobiles made up fully of parts listed in manufacturer specifications for each model was held in Charlotte, North Carolina. The sport (this label is often debated among philosophers of athletics) has expanded its fan base from a regional blue-collar group in the southeastern U.S. to a diverse nationwide audience in recent years.

The Winston Cup series is currently made up of 36 races per year, with 43 cars competing in each race (both of these numbers have varied over the years). The series generates a rich set of data and possibilities for comparisons
among drivers/crews. Students throughout the country have been exposed to NASCAR through national telecasts of races, as well as many promotional activities among the drivers. Interested students could “mine” the data and come up with many questions to answer as well as many ways to graphically describe the data.

The most important outcomes of each race is the driver’s finishing position and prize winnings. The first driver to cross the start/finish line on the last lap of the race wins the race. Once the winner has crossed the finish line (received the checkered flag), all other drivers remaining on the track complete their current lap. Second place goes to the second car on the “lead lap” to cross the finish line. Thus, all cars on the lead lap who cross the finish line will have completed the maximum number of laps. The finishing positions are ordinal outcomes as opposed to quantitative outcomes. It is not uncommon for over half the cars to finish within a lap or two of each other while several cars may only complete a handful of laps. Also, in terms of prize money, drivers who may finish within a lap of each other may receive vastly different prizes. Students typically are exposed quickly in any introductory course to the concept of interval/ratio scaled outcomes (prize money) and ordinal measures (finish position). They may be shocked to observe the differences in prize money among drivers who finish within very short distances of one another in a long race.

The fact that the same driver/crew teams participate each week (for the most part) allows for some interesting statistical questions to be posed. Unlike golf, where many of the top competitors play selective schedules, virtually all top teams compete in each race of the season. Throughout this paper, we will simply identify the driver/crew team by the driver.

Instructors may assign their students many different events of which to obtain (empirical) probabilities, as well as give them the potential to directly think through the relevant steps for obtaining conditional probabilities of interest. Further, students may be exposed to many applications of Bayes’ Rule, and obtain various probability distributions. An in-lab computer project is supplied that lets students obtain various probabilities by sifting through raw data with a spreadsheet. Similar projects could be used to obtain any number of descriptive statistics or graphical displays. Students may use statistical methods to explore the growth of NASCAR’s popularity based on payouts to drivers, make head-to-head comparisons of drivers’ racing skills, as well as to compute and compare rank correlation coefficients between starting and finishing positions. The racing skills of pairs of drivers can be compared by using simple methods of categorical data analysis for matched pairs, such as McNemar’s Test (McNemar, 1947). While many students may not have been exposed to this method, they typically have compared pairs of means and proportions for independent samples, and means based on paired samples; this gives instructors a means to fill in a void, using a very simple statistical test applied to real data. By the nature of the ranking of the starting and finishing positions, students can compare two rank correlation coefficients: Spearman’s $\rho$ (Spearman, 1904) and Kendall’s $\tau$ (Kendall, 1938). While most methods courses describe Spearman’s measure, many students have not been exposed to Kendall’s method, and this gives them an opportunity to compare two “competing” measures. The datasets afford instructors and students a wide range of possibilities to apply methods of descriptive and inferential statistics over various levels of sophistication (all of which are becoming more accessible to students, at least conceptually and computationally).

The driver dataset nascard.dat.txt contains the finishing and starting positions for each driver in 898 Winston Cup races between 1975 and 2003. Also included are the driver’s name, prize winnings for that race, number of laps completed, and car make. The dataset contains 34,884 observations at the driver level. Note that prize winnings are not necessarily monotone decreasing in finish position.

A second dataset nasarr.dat.txt contains race specific characteristics. It contains the series race (1,…,898), year (1975-2003), race within the year, number of cars, total payout, Spearman’s $\rho$, Kendall’s $\tau$, track length, laps completed, road track indicator, number of caution periods, number of lead changes, time to complete race, Consumer Price Index (CPI-U) for the month of the race, latitude/longitude coordinates, and track name. Note that the race distance can be obtained by multiplying the number of laps by the track length (this allows for races that were shortened due to weather). Also, average speed for the winning driver can be obtained by dividing miles by completion time (minutes) and multiplying by 60 (minutes/hour).

Data were obtained from the NASCAR website (www.nascar.com), as well as racing-reference.com. Information is given on these websites regarding all Winston Cup races between 1975 and 2003, and beyond. Information regarding the tracks participating over this period was obtained from these websites, and web searches for
information on tracks not currently participating in Winston Cup racing. Due to rapidly changing corporate sponsorship, race names are not included in the datasets.

2. Probability Exercises and Statistical Applications

In this section, we describe some activities that involve obtaining event probabilities and statistical analyses that instructors and students can apply with these datasets. The first subsection describes some opportunities to obtain various probabilities; we have run this exercise in a computer lab with a small honors section of an Introduction to Statistics class with 24 students working in pairs. The second subsection involves measuring and describing the growth of NASCAR’s popularity based on total prize money by year. The third subsection describes an analysis of matched pairs for categorical responses. The final subsection involves computation of rank correlation coefficients and tests of independence between starting and finishing positions.

2.1 Computing Basic Probabilities

Students can estimate probabilities of any number of events, as well as probability distributions. They can also be challenged to come up with events of interest, and then estimate their corresponding probabilities. Also, since many students struggle to understand the notions of conditional probabilities, any number of exercises can be set up to help them understand conditioning. It also allows them to think through the process of subsetting datasets to obtain the requested probabilities. Bayes’ Rule can be applied as well, and for students of mathematical statistics, the concept of Bayesian updating and posterior distributions can be applied to real world data. Some potential examples include:

- The probability a Ford wins the race
- The probability a Ford wins the race, obtained separately by year
- The probability distribution of finishing positions for Dale Earnhardt
- The probability Ford outperforms Chevrolet (based on average winnings per car)
- The probability the driver starting first wins the race
- The probability the driver starting first finishes at or below position $s$
- The probability the winner of the race started at or below position $s$
- The posterior distribution for the probability Ford is “better” than Chevrolet at the end of a season, based on a uniform (U(0,1)) prior distribution at the beginning

Students may be asked to manage (sort, select special cases within, and/or create new variables from) the datasets to obtain the specified probabilities, or the instructor may prepare worksheets or a program to select the cases for the students to obtain relevant probabilities. We feel that a combination of students doing the operations and instructor pre-preparation of some datasets will maximize the opportunities to obtain a wide range of probabilities. We have recently attempted this in a small computer lab with 24 students working in pairs, but there is no reason it could not be given as home exercises in larger classes (assuming availability of software). The students’ thought processes on determining the order of sorting to get the appropriate conditional probabilities was interesting to observe. The Appendix contains the in-lab project assignment/instructions.

2.2 Describing Growth in NASCAR Popularity (1975-2003)

Annual total payout to Winston Cup drivers are given in Table 1 and displayed in Figure 1. Actual payout, percent change, and adjusted payout are given. Adjusted payouts are based on the consumer price index for all urban consumers (CPI-U), obtained from the Bureau of Labor Statistics website. Their basis level is 1982-1984, so adjusted dollars are based on those levels. Based on these series, students can quantify numerically and describe graphically the rapid growth of the popularity of NASCAR, particularly in the early 1990s and beyond. Further, they can be asked to obtain percent change in adjusted dollars and compare those with the percent change in actual dollars.

| Table 1. Annual payouts (actual and adjusted to 1982-1984) for NASCAR Winston Cup Races ($1,000,000s) |
| Year | Payout ($1Ms) | % Change | Adjusted Payout ($1Ms) |
|------|--------------|----------|------------------------|
| 1975 | 2.41         | --       | 4.48                   |
| 1976 | 3.42         | +41.9    | 6.01                   |
| 1977 | 3.59         | +5.0     | 5.92                   |
| 1978 | 3.94         | +9.7     | 6.04                   |
| 1979 | 4.92         | +24.9    | 6.78                   |
| 1980 | 5.41         | +10.0    | 6.57                   |
| 1981 | 6.01         | +11.1    | 6.61                   |
| 1982 | 7.08         | +17.8    | 7.34                   |
| 1983 | 7.40         | +4.5     | 7.43                   |
| 1984 | 8.58         | +15.9    | 8.26                   |
| 1985 | 9.06         | +5.6     | 8.42                   |
| 1986 | 10.40        | +14.8    | 9.49                   |
| 1987 | 10.91        | +4.9     | 9.60                   |
| 1988 | 11.88        | +8.9     | 10.04                  |
| 1989 | 12.83        | +8.0     | 10.35                  |
| 1990 | 14.36        | +11.9    | 10.99                  |
| 1991 | 15.46        | +7.7     | 11.35                  |
| 1992 | 18.21        | +17.8    | 12.98                  |
| 1993 | 21.06        | +15.7    | 14.57                  |
| 1994 | 26.32        | +25.0    | 17.76                  |
| 1995 | 33.34        | +26.7    | 21.88                  |
| 1996 | 38.00        | +14.0    | 24.22                  |
| 1997 | 48.68        | +28.1    | 30.33                  |
| 1998 | 66.50        | +36.6    | 40.80                  |
| 1999 | 77.43        | +16.4    | 46.48                  |
| 2000 | 85.37        | +10.3    | 49.58                  |
| 2001 | 117.86       | +38.1    | 66.55                  |
| 2002 | 129.12       | +9.6     | 71.77                  |
| 2003 | 143.7        | +11.3    | 78.10                  |
Table 1 and Figure 1 depict the rapid growth in popularity of NASCAR over the past 30 years as measured by total payout. An average annual growth rate can be computed from the values in the % change column by taking the geometric mean of the growth rates, where the growth rate is computed from the multiplier: $1 + G_i = (1 + (\% \text{ change}/100))$ for each year:

$$\bar{G} = \sqrt[28]{\prod_{i=1976}^{2002} (1 + G_i)} - 1 = 0.1573$$

Thus the average annual growth rate (obtained in the manner that an average rate of return is computed in finance) is 15.7%. Students could be asked to compute these for adjusted dollars, or different races, or for individual drivers. Also, students can compare the geometric mean with the arithmetic mean or the median, which are less appropriate for describing growth rates.

Students can estimate trend lines for the payouts, assuming linear and exponential growth models and compare their fits (and sadly, may be disappointed). Correlation and regression are being taught earlier in introductory statistics courses, and many students are now being exposed to these methods prior to basic probability (e.g. Moore and McCabe, 2006). Most statistical software packages have options to fit these models. They may be asked to conceptually describe the relationship since it doesn’t appear to fit well to either model which place severe restrictions on growth (a combination of the two seems to fit well visually).

### 2.3 Comparing Pairs of Drivers

Drivers have very loyal fan bases, much as teams do in sports such as football or baseball, as attested to widespread marketing of drivers and the prevalence of NASCAR team paraphernalia. Due to the fact that most top drivers compete in virtually every race, and generally have fairly long careers, students have ample opportunity to compare pairs of drivers in head-to-head competition. Due to the ordinal nature of the start and finish outcomes, we can exploit a simple method for comparing matched pairs of categorical outcomes.

While virtually every statistics textbook covers comparisons of two means and proportions for independent samples and comparisons of means for paired samples, most do not fill in an obvious hole: comparisons of
proportions for paired samples. A very simple procedure can be used to test for differences in proportions (McNemar, 1947). We describe the test and confidence interval that instructors can easily introduce to their students through this data.

For any pair of drivers (say A and B), we have a starting order (A ahead of B or B ahead of A) and a finishing order. If Driver A starts and finishes ahead of (or behind) B, then they completed the race in the same order. Because starting order generally represents the cars’ levels of performance for that weekend, we can’t say anything about the two drivers’ relative performances based on starting position. However, if Driver A starts ahead of B, and B finishes ahead of A, we might surmise that Driver B outperformed A in that race (or at least covered more ground). Likewise, if A started behind, but finished ahead of B, we could say A outperformed B.

Students can conduct a test to compare proportions based on matched pairs (see Agresti, 2002, Chapter 10 or Agresti, 1996, Chapter 9). The basic idea is to set up a 2x2 table with the driver who started the race ahead forming rows and the driver who finished ahead in the columns. Table 2 shows the general form and notation.

| Start | Finish | A ahead | B ahead | Total |
|-------|--------|---------|---------|-------|
| A ahead |       | n_{AA}  | n_{AB}  | n_{A+}|
| B ahead |       | n_{BA}  | n_{BB}  | n_{B+}|
| Total  |       | n_{+A}  | n_{+B}  | n_{++}|

Table 2. Cross-classification table for pairs of drivers’ start and finish ordering

Students can test whether the two drivers’ race abilities differ, where is the (true) probability that A starts ahead of B and is the probability that A finishes ahead. Defining \( \delta = \pi_{+A} - \pi_{A+} \), we could say that the two drivers’ racing skills are equal if \( \delta = 0 \), that is, the probability that driver A beats B is equal to the probability that driver A starts ahead of B. If \( \delta > 0 \), then driver A tends to outperform B on the track; if \( \delta < 0 \), B outperforms A.

The following statistic can be used to test whether \( \delta = 0 \) (see Agresti, 2002, p. 411 or Agresti, 1996, p. 228):

\[
Z_0 = \frac{n_{BA} - n_{AB}}{\sqrt{n_{BA} + n_{AB}}}
\]

(1)

This test statistic is the signed square root of McNemar’s chi-square statistic (McNemar, 1947). For large samples, this statistic is approximately normal. An exact test can be conducted based on the binomial distribution, where is distributed Binomial with \( n = n_{BA} + n_{AB} \) and \( p = 0.5 \) under the hypothesis of no driver skill difference.

Based on the normal approximation, values of \( Z_0 \) above \( Z_{\alpha/2} \) are evidence in favor of A being the better of the two drivers in racing conditions, values less than \( -Z_{\alpha/2} \) provide evidence that B is better.

This allows for students to make use of multiple comparisons as well. Suppose they would like to make pairwise head-to-head comparisons among \( k \) drivers. Then, they can see they will be making pairwise comparisons among
\( C = k(k - 1)/2 \) pairs of drivers. If they wish to keep the experimentwise error rate at level \( \alpha \), they can use Bonferroni’s (conservative) method, and make each individual comparison at \( \alpha^* = \alpha / C \).

We demonstrate by making pairwise comparisons among the following set of drivers: Dale Earnhardt (Sr.), Jeff Gordon, Darrell Waltrip, Terry Labonte, and Bill Elliott. Students can be assigned different pairs of drivers, or choose pairs of drivers they are familiar with, and be asked to conduct the test for their pair(s). First, students must obtain datasets containing all races for each driver, then merge (side-by-side) the datasets for each pair by race, including only races that both drivers competed in. Also, note that start and finish variables must be labeled differently for the 2 drivers (e.g. startde, finishde, startjg, and finishjg when comparing Dale Earnhardt and Jeff Gordon). This gives students a challenging problem in managing and combining large datasets (without the risk of permanently damaging or destroying them). Table 3 gives the results for all \( C = 10 \) pairs of drivers. The critical value, based on Bonferroni’s method, with \( \alpha = 0.05 \) is \( z_{0.025} = 2.81 \) Also included are simultaneous 95% confidence intervals for the differences, \( \delta \). The estimate \( d \) of \( \delta \) and its estimated standard error can be computed as (see e.g. Agresti, 2002, pp. 410-411 or Agresti, 1996, pp. 227-229, although the notation for standard error is given in different forms):

\[
d = p_{A*} - p_{A+} \quad \delta = \sqrt{\left( p_{AB} + p_{BA} \right) - \left( p_{AB} - p_{BA} \right)} \quad p_{\hat{q}} = \frac{n_{\hat{q}}}{n_{11}}
\]

Note that the standard error for the confidence interval does not place the constraint that the true proportions are equal, and is more complicated than that for the test. Students may be asked how this is analogous to the case for independent samples.

We make the following conclusions (with an experimentwise Type I error rate of 0.05):

- Dale Earnhardt is a tougher in race driver than Jeff Gordon and Bill Elliott (\( \delta > 0 \))
- Jeff Gordon is a weaker in race driver than Darrell Waltrip and Terry Labonte (\( \delta < 0 \))
- Terry Labonte is a tougher in race driver than Bill Elliott (\( \delta > 0 \))

We can summarize the results by ordering the drivers and joining pairs of drivers who do not differ significantly with lines.

\[
\begin{array}{cccc}
JG & BE & DW & TL & DE \\
\end{array}
\]

### Table 3. Observed frequency of start/finish orderings for 10 pairs of drivers and simultaneous 95% Confidence Intervals for (\( \delta \))

| Driver A       | Driver B | \( n_{AA} \) | \( n_{AB} \) | \( n_{BA} \) | \( n_{BB} \) | \( z_0 \) | 95% CI for (\( \delta \)) |
|----------------|----------|---------------|---------------|---------------|---------------|-----------|---------------------------|
| Earnhardt      | Gordon   | 41            | 31            | 82            | 104           | 4.80      | (0.087, 0.308)            |
| Earnhardt      | Waltrip  | 231           | 112           | 152           | 161           | 2.46      | (-0.008, 0.130)           |
| Earnhardt      | Labonte  | 230           | 128           | 155           | 154           | 1.60      | (-0.030, 0.111)           |
| Earnhardt      | Elliott  | 188           | 91            | 171           | 149           | 4.94      | (0.059, 0.208)            |
| Gordon         | Waltrip  | 171           | 48            | 18            | 7             | -3.69     | (-0.214, -0.032)          |
2.4 Computing Spearman’s $\rho$ and Kendall’s $\tau$

Spearman’s $\rho$ is a measure of correlation between two variables based on the relative ranks of each observation (Spearman 1904). Intuitively, if we have a set of $n$ pairs $(X_i, Y_i)$, we replace the data pairs with their ranks, and compute Pearson’s product moment correlation coefficient based on the ranks. Note that the mean rank is $(n + 1)/2$, where $n$ is the number of drivers. For the current data, students can compute Spearman’s $\rho$ separately for each race, measuring the association between starting and finishing positions which are ranks. Thus, for a race with $n$ drivers, and starting and finishing pairs $(S_i, F_i)$, we get:

\[
\hat{\rho} = \frac{\sum_{i=1}^{n} \left( S_i - \frac{n+1}{2} \right) \left( F_i - \frac{n+1}{2} \right)}{\sqrt{\sum_{i=1}^{n} \left( S_i - \frac{n+1}{2} \right)^2 \sum_{i=1}^{n} \left( F_i - \frac{n+1}{2} \right)^2}}
\]

We treat this as a random variable in the sense that today’s race is one realization of a conceptual population of races that could have been run. This quantity has been computed in the nascarr.dat.txt dataset, but can be directly computed from the full dataset nascard.dat.txt. Students can compute this statistic on their own and also observe the empirical distribution of this statistic in repeated samples. Further, students may try to “explain” the variation in this measure (and Kendall’s $\tau$ below) by fitting a regression model, relating the correlation measure(s) to: track length, numbers of laps, caution flags, lead changes, and drivers. Students can be challenged by asking that if their goal was predicting the measure prior to the race beginning, which of these predictors should be used in the model. They may also compare the fits of the two models.

Kendall’s $\tau$ has also been computed for each race. For a given race, there are $n(n - 1)/2$ pairs of drivers. Beginning with the driver who finished first, we count how many drivers started behind him/her, then we proceed to driver 2, and see how many drivers that finished behind him/her started behind him/her and so on (Kendall 1938, Kendall and Gibbons, 1990). The total count will be called $k$. Thus, if a driver who won the race had started first, (s)he would contribute $n - 1$ to $k$, while if a driver who won had started last, (s)he would contribute 0 to $k$. Then, for a race with $n$ drivers, we have:

\[
\hat{\tau} = \frac{4k}{n(n-1)} - 1
\]

Note that if the drivers end in the exact order they start, $k = (n - 1) + (n - 2) + ... + 1 = (n - 1)n/2$ and Kendall’s $\tau$ takes on the value 1, similarly, if drivers perfectly reversed their order it would take on $-1$.

Tests of independence between starting and finishing position can be conducted based on both Spearman’s $\rho$ and Kendall’s $\tau$. The test statistics (based on no ties among the starting or finishing positions) are:

| Gordon | Labonte | 218 | 30 | 35 | -4.77 | (-0.216, -0.059) |
|--------|---------|-----|----|----|-------|-----------------|
| Gordon | Elliott | 183 | 69 | 65 | 38    | -0.35           |
|        |         | 124 | 38 | 182| -1.16| (-0.099, 0.041) |
| Waltrip| Labonte | 205 | 143| 124| 182   | 1.13            |
|        |         | 121 | 217| 238| 1.13  | (-0.042, 0.098) |
| Labonte| Elliott | 162 | 123| 181| 238   | 3.33            |
|        |         |     |    |     |       | (0.013, 0.151)  |
Both statistics are approximately standard normal for large samples when there is no association. If we use these to test for each race whether there is a positive association between starting and finishing position, based on \( \alpha = 0.05 \) significance level (concluding there is a positive association if \( Z \geq 1.645 \)) we obtain the following results in Table 4.

|                | Kendall                              |
|----------------|--------------------------------------|
|                | Positive Association \( Z_\tau \geq 1.645 \) | No Association \( Z_\tau < 1.645 \) |
| Positive Association \( Z_\rho \geq 1.645 \) | 637 (70.9%) | 12 (1.3%) |
| No Association \( Z_\rho < 1.645 \) | 12 (1.3%) | 237 (26.4%) |

Thus, they virtually always give the same conclusion regarding association between starting and finishing positions. Note that students could apply McNemar’s test here to determine whether one measure is more/less likely to conclude there is a positive association than the other.

A plot of Spearman’s \( \rho \) and Kendall’s \( \tau \) across time is given in Figure 2, where we combine the measures over each year, treating races as blocks (Taylor, 1987). We average the measures over each year with weights equal to the number of cars in the race. Note that the level of correlation between starting and finishing positions appears to have dropped off quite a bit since the mid 1990s, possibly due to increased competition among teams and more money being spent on equipment as the payouts have grown. Students may think of alternative explanations of this and further investigate it, as many rules changes and changes in equipment have been made over the years.
As a result, we have 898 pairs \((\hat{\rho}_i, \hat{\tau}_i)\). Students can compute Pearson’s product moment coefficient of correlation as (where \(\bar{\rho}\) and \(\bar{\tau}\) are the sample means for each measure):

\[
r = \frac{\sum_{i=1}^{898} (\hat{\rho}_i - \bar{\rho})(\hat{\tau}_i - \bar{\tau})}{\sqrt{\sum_{i=1}^{898} (\hat{\rho}_i - \bar{\rho})^2 \sum_{i=1}^{898} (\hat{\tau}_i - \bar{\tau})^2}}
\]

For this series, the correlation coefficient is \(r = .9908\). Thus (not surprisingly) there is a strong correlation between these two rank correlation coefficients. A scatterplot of the rank correlations is given in Figure 3. While these measures are based on different criteria, their levels are very highly correlated across samples. Students could empirically obtain the sampling distribution of the correlation coefficient \(r\) when the correlation is high, by taking many random samples of races and observe its distribution of sample values.

### 3. Other Applications

The datasets could be used in many different ways to help teach statistical techniques to students at various levels of sophistication. Examples of other possibilities, beyond those described in the previous section, include (in approximately ascending order of complexity):

- Data management techniques:
  - Subsetting datasets: By year, or racetrack
Building case histories: Annual winnings and finishes by driver

- Plots:
  - Histogram of rank correlation measures or driver payouts in a year
  - Pie Charts of payout shares by car make by year
  - Box Plots of driver payouts in a year
  - Line Plots of median finish versus starting position (or any time series)
  - Scatterplots of lead changes by crashes or lead changes by race length
  - Median finish versus starting position
  - Kaplan-Meier survival curves for total laps completed in season by groups of drivers

- Numerical Descriptive Measures
  - Average winnings for a driver in a season
  - Proportion of drivers completing all laps in a race
  - Point pattern analysis: The mean centre and standard distance of race locations (either weighted by payout or unweighted) can be obtained by year, showing the location and spread of popularity of the sport over time (see e.g. Fotheringham, Brunsdon, and Charlton 2000, p. 136)

\[
\text{Mean Center} : (\hat{\mu}_x, \hat{\mu}_y) = \left( \frac{\sum w_i x_i}{\sum w_i}, \frac{\sum w_i y_i}{\sum w_i} \right)
\]

\[
\text{Standardized Distance} : d_s^2 = \frac{\sum w_i \left( (x_i - \hat{\mu}_x)^2 + (y_i - \hat{\mu}_y)^2 \right)}{\sum w_i}
\]

- Measuring competition: Economic measures such as the Concentration Ratio or the Herfindahl-Hirschman Index (HHI) can be applied to drivers winnings to determine how competition has changed over time. For a season with \( N \) drivers, where \( s_i \) is the percent of total prize money won by the \( i \)th winningest driver in prize share, these measures are (see e.g. Mansfield, 1999):

\[
\text{Concentration Ratio} : CR_k = s_1 + \cdots + s_k \quad (k = 1, \ldots, N)
\]

\[
\text{Herfindahl - Hirschmann Index} : HHI = s_1^2 + \cdots + s_N^2
\]

- Extreme measurements: Determine the fraction of potential winnings for each driver each season (where total potential is the sum of payouts to winning driver). Obtain the maximum winning fraction for each season and obtain its distribution, and plot over time.

- Comparison of car makes: Fans are very loyal to manufacturers. Ford and Chevrolet could be compared with respect to average winnings in constant dollars, possibly being compared separately by year. Nonparametric methods could be used to compare all makes within races (Wilcoxon Rank-Sum or Mann-Whitney Tests)

- Regression modeling: Try and “explain” the variation in rank correlations, applying model building techniques to fit a “best” model. Predictor variables could include: Track length, number of laps, number of caution periods, and number of lead changes.

- Time series applications: Rank correlations or car make winning shares could be modeled over time with “interventions” at the end of each season representing rules changes and re-tooling of cars, as well as new competitors.

- Survival Analysis: Kaplan-Meier estimates of driver “survival” through races by estimating the survival function with respect to lap completion by car make.

- Generalized Linear Models:
  - Logistic Regression: Determining the probability that Driver A beats B, as a function of the difference in their starting positions, or the probability the driver starting first finishes in top 5 as a function of laps and track length.
  - Poisson Regression: Modeling the number of caution periods as a function of number of laps, track length, and number of drivers.
  - Negative Binomial Regression: Modeling the number of lead changes as a function of number of laps, track length, and number of drivers (we have found that the Poisson model does not fit well (over
1975-1979), but the Negative Binomial does).
- Application of the logistic-normal distribution for continuous proportions: Within years, we have a
  series of proportions of prize money going to each car make for each race. These represent continuous
  proportions (see Agresti, 2002, p. 265, Problem 6.33 for a description), and might be modeled as a
  function of track length and number of laps.

4. Conclusions

In this paper, we have introduced datasets containing results from all NASCAR races from 1975-2003 inclusive at
the driver and race levels. Examples have been chosen to demonstrate activities for students that involve: obtaining
basic and conditional probabilities; describing growth in payouts in real terms, percent changes, and adjusted
terms; learning to conduct a simple test for proportions based on paired samples; and making use of the ordinality
of start and finish positions to work with two measures of rank correlation. A series of other potential applications
is also offered to instructors and students. We feel with the growing popularity of NASCAR among both males and
females, these datasets would be of interest to statistics, economics, and math instructors and their students.

5. Getting the Data

The file nascard.dat.txt is a text file containing 34884 rows. Each row corresponds to a particular driver competing
in a particular race. The file nascard.txt is a documentation file describing the variables.

The file nascarr.dat.txt is a text file containing 898 rows. Each row corresponds to a particular race in a particular
year. The file nascarr.txt is a documentation file describing the variables.

Appendix 1A – Key to Variables in driver dataset nascard.dat.txt

| Columns | Variable                | Comments             |
|---------|-------------------------|----------------------|
| 1 - 3   | Series Race             | 1, 2, …,898          |
| 6 - 9   | Year                    | 1975, …, 2003        |
| 12 - 13 | Race/Year               | Format F2.0          |
| 16 - 17 | Finishing Position      | Format F2.0 (1=Winner)|
| 20 - 21 | Starting Position       | Format F2.0          |
| 24 - 26 | Laps Completed          | Format F3.0          |
| 29 - 35 | Winnings                | Format F7.0 (In dollars)|
| 38 - 39 | Number of cars in race  | Format F2.0          |
| 42 - 50 | Car Make                | String of Length 9   |
| 53 - 82 | Driver                  | String of Length 30  |

Appendix 1B – Key to Variables in race dataset nascarr.dat.txt

| Columns | Variable       | Comments             |
|---------|----------------|----------------------|
|         |                |                      |
Appendix 2

NASCAR In-Class Project – Probability

You have been hired to describe the race history of NASCAR Winston Cup races (1975-2003), by obtaining the following probabilities from your EXCEL worksheets provided in class. Please complete the following parts. (Included are first lines of prepared worksheet)

Worksheet 1: All races held from 1975-2003 (race level data):

| Race | Year | TrkLength | Laps | Cautions | Leadchng | RaceTime |
|------|------|-----------|------|----------|----------|----------|
| 1    | 1975 | 2.54      | 191  | 5        | 13       | 304.433  |
| 2    | 1975 | 2.50      | 200  | 3        | 19       | 195.250  |
| 3    | 1975 | 0.75      | 500  | 7        | 2        | 217.050  |

Probability there were less than or equal to 3 caution flags ________________
Probability there were more than 10 lead changes ________________
Probability there were at least twice as many lead changes as cautions ________________
Probability the average speed was over 150 miles per hour ________________
Probability the average speed was below 100 miles per hour ________________

Worksheet 2: All drivers participating in Daytona 500 races from 1975-2003

| Race | Finish | Start |
|------|--------|-------|
Probability the driver who started first finished first _______________________
Probability the driver who started first finished in top ten _______________________
Probability the driver who finished first started first _______________________
Probability the driver who finished first started in top ten _______________________

Worksheet 3: All Drivers starting first and second (side-by-side format)

| Race | TrkLength | Start1 | Finish1 | Car1  | Start2 | Start2 | Finish2 |
|------|-----------|--------|---------|-------|--------|--------|---------|
| 1    | 2.54      | 1      | 1       | Matador | 2      | 2      | Mercury |
| 2    | 2.50      | 1      | 28      | Chevrolet | 2      | 4      | Mercury |
| 3    | 0.75      | 1      | 1       | Dodge  | 2      | 3      | Chevrolet |

Probability that driver starting first beats driver starting second._______
Probability that the first driver beats the second given track length is \( \leq 1 \) mile _______
Probability that the first driver beats second given track length \( \geq 2.0 \) miles_________
Probability driver starting first drove a Ford _________________
Probability driver starting first drove a Chevy _________________

Worksheet 4: All Drivers finishing first and second (side-by-side format)

| Race | TrkLength | Laps  | Start1 | Finish1 | Car1  | Start2 | Start2 | Finish2 |
|------|-----------|-------|--------|---------|-------|--------|--------|---------|
| 1    | 2.54      | 191   | 1      | 1       | Matador | 2      | 2      | Mercury |
| 2    | 2.50      | 200   | 32     | 1       | Chevrolet | 3      | 2      | Matador |
| 3    | 0.75      | 500   | 1      | 1       | Dodge  | 3      | 2      | Chevrolet |

Probability that driver finishing first started ahead of driver finishing second._______
Probability that event described above occurred given race length \( \geq 350 \) miles _______
Probability driver finishing first drove a Ford _________________
Probability driver finishing first drove a Chevy _________________

Notes on variables:

Race Length (Miles) = Laps completed by winner x Track length

Completion Time is measured in minutes, divide by 60 to change to hours

Use these to compute speeds in miles per hour

References
Agresti, A. (1996), *An Introduction to Categorical Data Analysis*, New York: Wiley.

Agresti, A. (2002), *Categorical Data Analysis*, 2nd Ed., Hoboken, New Jersey: Wiley.

Fotheringham, A.S., Brunsdon, C., and Charlton, M. (2000), *Quantitative Geography*, London: Sage.

Kendall, M.G. (1938), “A New Measure of Rank Correlation,” *Biometrika*, 30, 81-93.

Kendall, M. and Gibbons, J.D. (1990), *Rank Correlation Methods*, 5th Ed., London: Edward Arnold.

Mansfield, E. (1999), *Managerial Economics*, 4th Ed., New York: W.W. Norton.

McNemar, Q. (1947), “Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages,” *Psychometrika*, 12, 153-157.

Moore, D.S. and McCabe, G.P. (2006), *Introduction to the Practice of Statistics*, 5th Ed., New York: W.H. Freeman.

*NASCAR Record & Fact Book* (2004 Ed.), St. Louis, MO.: Sporting News Books.

Spearman, C. (1904), “The Proof and Measurement of Association Between Two Things,” *American Journal of Psychology*, 15, 72-101.

Taylor, J.M.G. (1987), “Kendall’s and Spearman’s Correlation Coefficients in the Presence of a Blocking Variable,” *Biometrics*, 43, 409-416.

Larry Winner
Department of Statistics
University of Florida
Gainesville, FL 32611-8545
U.S.A.
winner@stat.ufl.edu