Comparing League Strength via the Soccer Power Index

Andrew Burkard

August 2, 2018

Abstract

We examine FiveThirtyEight's Soccer Power Index (SPI) to compare the quality of top European soccer leagues. Using non-parametric statistical methods, we determine that there is not a significant difference in the SPI between clubs in the first division of the English, Spanish, German, and Italian first divisions. We also determine that the SPI ratings of the average clubs in these leagues did not change significantly over the past season. Finally, we determine that the average English Football League Championship team has a higher SPI rating than the average Major League Soccer team.

1 Introduction

European club soccer is a high-stakes, high-earning business, with top clubs regularly pulling in hundreds of millions of dollars in annual revenue [del18]. A subject of great debate is which league and clubs represent the highest level of play. One system that tries to answer this question is the Soccer Power Index (SPI) by FiveThirtyEight. While its particular implementation is somewhat of a black box, the SPI uses a combination of play-by-play and transfer market data to rate clubs across 27 leagues on a common scale [Boi17b]. One intriguing finding in a previous edition of the SPI is that there are 4 leagues the authors consider to be "Tier 1": the English Premier League, Germany's Bundesliga, Spain's La Liga, and Italian Serie A [Boi17a]. We attempt to confirm whether this is still the case by applying non-parametric statistical tests to the SPI ratings of the individual clubs from these four leagues. Additionally, we use SPI along with appropriate tests to determine whether there were changes in the relative strength between these leagues over the course of the 2017-18 season. Finally, we go inter-continental to compare the strengths of two weaker leagues: Major League Soccer and England's second division.

For all of these questions, we utilize suitable non-parametric statistical tests. While we could rely on t-tests or other parametric methods, there are various reasons why we should prefer non-parametric methods. In terms of power, which is the probability of correctly rejecting a false null hypothesis, the t-test is only optimal when the underlying population distribution of the data is approximately normal. Additionally, for a t-test to be optimal, it is required that the variance within treatment groups be identical. These properties highlight a key distinction of nonparametric from traditional parametric analytical methods: We do not have to assume the data are drawn from a given parameterized probability distribution to perform nonparametric tests [Hig04]. As such, we can report robust results that are less likely to fall victim to faulty or violated assumptions.

2 Data

We were able to gather SPI data for all matches in 27 leagues during the 2016-17 and 2017-18 seasons. Each record contains the SPI ratings of both teams going into the match, as well as various predictive attributes based on SPI such as expected goals and win probabilities. Since we are only interested in the SPI ratings themselves, we ignore these features.

To determine the 2017-18 start-of-season SPI ratings, which are needed to answer our second question, we use the pre-match SPI ratings for each club's first match in August 2017, the month in which all of the examined leagues begin. For the end-of-season ratings, we simply query a separate table containing the current ratings, as they have not been updated to include summer activity.

3 Methods

3.1 Is there a difference between Tier 1 leagues?

To determine whether there is a difference between any of the four Tier 1 leagues (Germany, Spain, England, Italy), we run a permutation F-test with four groups, where each contains the current ratings of the clubs in the respective league.

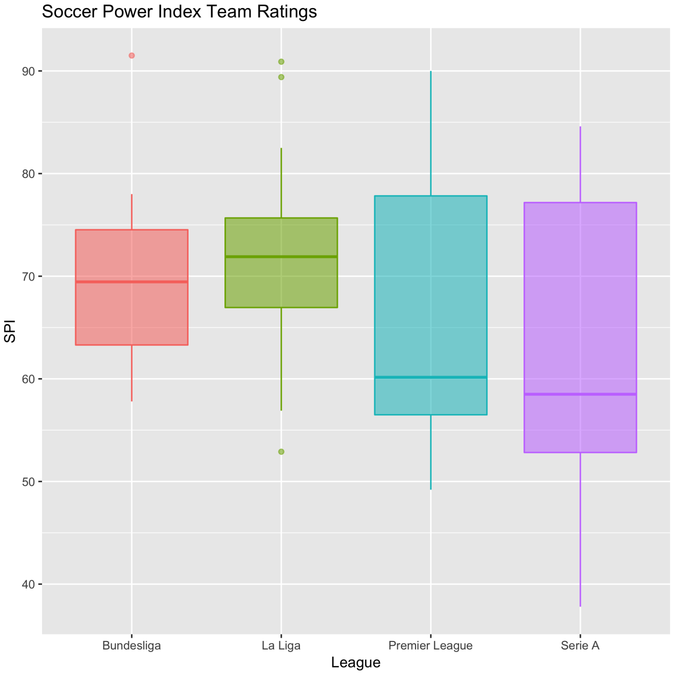

Figure 1: Distribution of the SPI ratings for clubs in different leagues after the 2017-18 season

We see in figure 1 that Germany and Spain appear have higher medians than England and Italy, but the latter have greater variance. However, cannot say whether this is statistically significant until running the test. We use the null hypothesis $H_0 : F_1(x) = ... = F_4(x)$, that is, the SPI ratings of the clubs within each league have the same distribution. For the alternative hypothesis, we have $H_A : F_i(x) \leq F_j(x)$ or $F_i(x) \geq F_j(x)$ for at least one pair (i, j) with strict inequality holding for at least one x. The permutation F-test assumes exchangeability under the null hypothesis. Note that because there are a total of 78 clubs across the 4 leagues (split 18-20-20-20), we must random sample the permutation distribution for computational feasibility.

3.2 Did League strength change over 2017-18?

To compare the strength of a league before and after the season, we look to the change in SPI ratings among its constituent clubs. However, because the pre-season SPI of an individual club is certainly correlated with its post-season SPI, we cannot rely on the i.i.d. assumption necessary for a basic permutation test. Instead, we run a paired-comparison permutation on the pre-season and post-season SPI ratings. We do this separately for each of the four leagues. Let $F()$ be the CDF of the differences in SPI ratings, D_i. In each case, the null hypothesis is $H_0 : F(x) = 1-F(x)$, i.e. F is symmetric because teams are equally likely to improve or regress. The alternative hypothesis is that the SPI of teams within a league shifted by some non-zero amount: $H_A : F(x) = G(x - \theta), \theta \neq 0$.

3.3 MLS vs. EFL Championship

Finally, seek to answer whether the second division of English Football (the confusingly named Championship) is actually superior to the highest division of U.S. Soccer, Major League Soccer (MLS). We do this via simple two sample permutation tests for the mean and median using the end-of-season 2018 ratings. The null hypothesis will be that the two distributions are equal,

$H_0 : F_1(x) = F_2(x)$, while the alternative is that the Championship is better according to SPI, $H_A : F_1(x) \geq F_2(x)$ with strict inequality for at least one x.

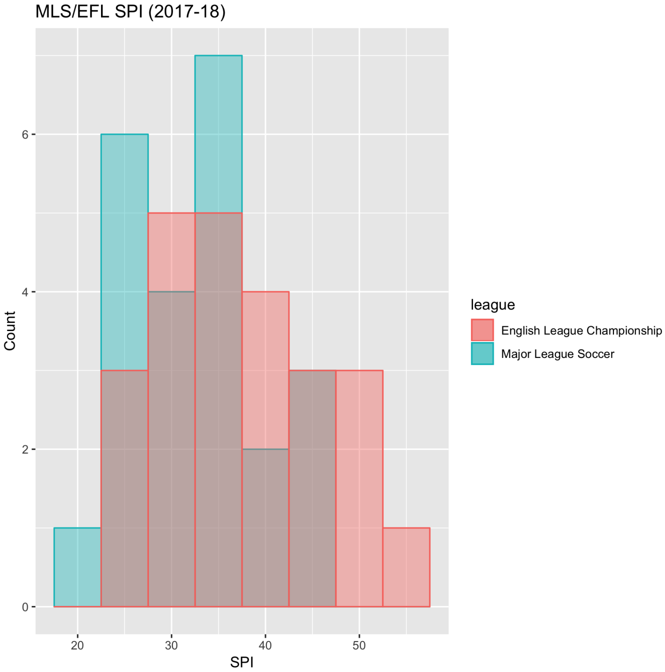

Figure 2: Distribution of the SPI ratings for EFL/MLS clubs after the 2017-18 season

Because the permutation requires exchangeability under the null hypothesis, and we do not know whether the means of the two distributions are identical, we must first determine whether the variances are different via a permutation test for deviances. In figure 2 it appears there may be greater deviation in the Championship than in MLS, perhaps owing to the fact that its top teams will be promoted to the Premiership. For this test, the null hypothesis is that the deviance of the two distributions are equal, $H_0 : \sigma_1 = \sigma_2$. Because we utilize a two-sided test, the alternative hypothesis is related to the ratio of absolute mean deviances, $H_A : \frac{\max(\sigma_1,\sigma_2)}{\min(\sigma_1,\sigma_2)} > 1$.

4 Results

4.1 Which league is best?

After running 10,000 samples of the permutation distribution, the four-way permutation F-test gives an test statistic of $F = 2.251$, resulting in a p-value of $\boldsymbol{p = .0995}$. Because this is not significant at the $\alpha = .05$ level, we fail to reject the null hypothesis H_0. As such, we cannot conclude that there is any difference in quality among the four Tier 1 leagues as determined by SPI. Because of this, there is no need to do any multiple comparison tests between leagues.

4.2 Did League strength change over 2017-18?

We again run 10,000 random samples for each paired comparison permutation test for each league. The resulting mean differences and p-values are reported in table 4.2.

League	$\bar{d}$	p
Bundesliga	-0.7233	.6158
La Liga	0.2490	.8670
Premier League	0.6770	.5811
Serie A	-0.4855	.6297

None of the permutation tests produce a significant p-value at the $\alpha = .05$ level. Therefore we fail to reject the null hypothesis in each case. We cannot conclude for any of the four leagues that there was a significant change in its overall quality over the 2017-18 season. Had we found a significant p-value in one of the tests, we would have needed to be careful to use an appropriate adjustment, as running multiple tests can make it more likely we produce "significant" p-values.

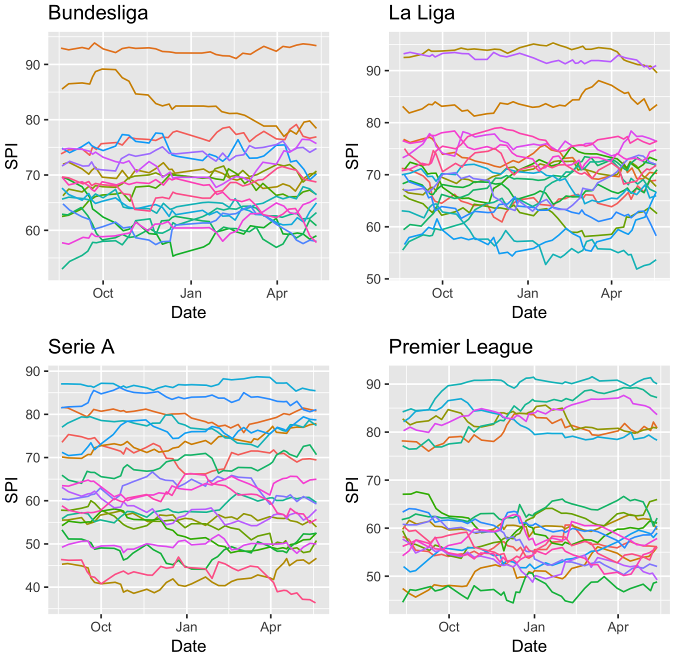

Figure 3: Changes in SPI of clubs over the 2017-18 season. Notice that the overall trend tends to be steady for all leagues, despite individual clubs having differing fortunes.

4.3 MLS vs. EFL Championship

The permutation for deviances produces a p-value of $\boldsymbol{p = .1548}$, leading us to fail to reject the null hypothesis that the two leagues SPI distributions have the same variance. Because of this result, we have satisfied the assumptions necessary to carry on with the permutation tests for the difference between means and medians.

The permutation test for the differences between medians gives a p-value of $\boldsymbol{p = .078}$, causing us to fail to reject the null hypothesis that there is no difference between the median SPI rating of the distribution for each league. However, the permutation test for differences between means does return a significant p-value of $\boldsymbol{p = .01}$. Thus, we can conclude that the mean average EFL Championship club is better than the mean average MLS team.

5 Discussion

We found that there is no statistically significant difference between the top four leagues in the SPI ratings. Perhaps this is unsurprising, given the creators of SPI make this proclamation when introducing the system. Even so, it runs counter to the conventional wisdom of the UEFA Association Club Coefficients, which rates La Liga (106.998) significantly above the Premier League (79.605), Serie A (76.249), and the Bundesliga (71.427) [UEF18]. It's possible this discrepancy is due to UEFA's consideration of only teams competing in the Champions League and Europa League competitions, which have been dominated by Spanish teams Real Madrid, FC Barcelona, and Sevilla in recent years. The power of the Barcelona-Real Madrid duopoly can also be seen in SPI in 3.

We should be careful to note that any conclusions reached in this research are wholly dependent on the robustness of the Soccer Power Index model, and are not necessarily indicative of the true quality of teams or leagues themselves. That is, while we can definitively conclude that the average EFL Championship club has a higher SPI than the average Major League Soccer team, it takes another logical jump to conclude that the average Championship club is "better" than the average

MLS team. Regardless, the questions we examined above are certain to be part of a continuing debate.

References

[Boi17a] Jay Boice. How our club soccer projections work, Jan 2017.

[Boi17b] Jay Boice. What's new in our 2017-18 club soccer predictions, Aug 2017.

[del18] Deloitte football money league 2018, 2018.

[Hig04] J.J. Higgins. *An Introduction to Modern Nonparametric Statistics*. Duxbury advanced series. Brooks/Cole, 2004.

[UEF18] UEFA.com. Member associations - uefa coefficients - country coefficients, Jul 2018.

Appendix A Additional Figures

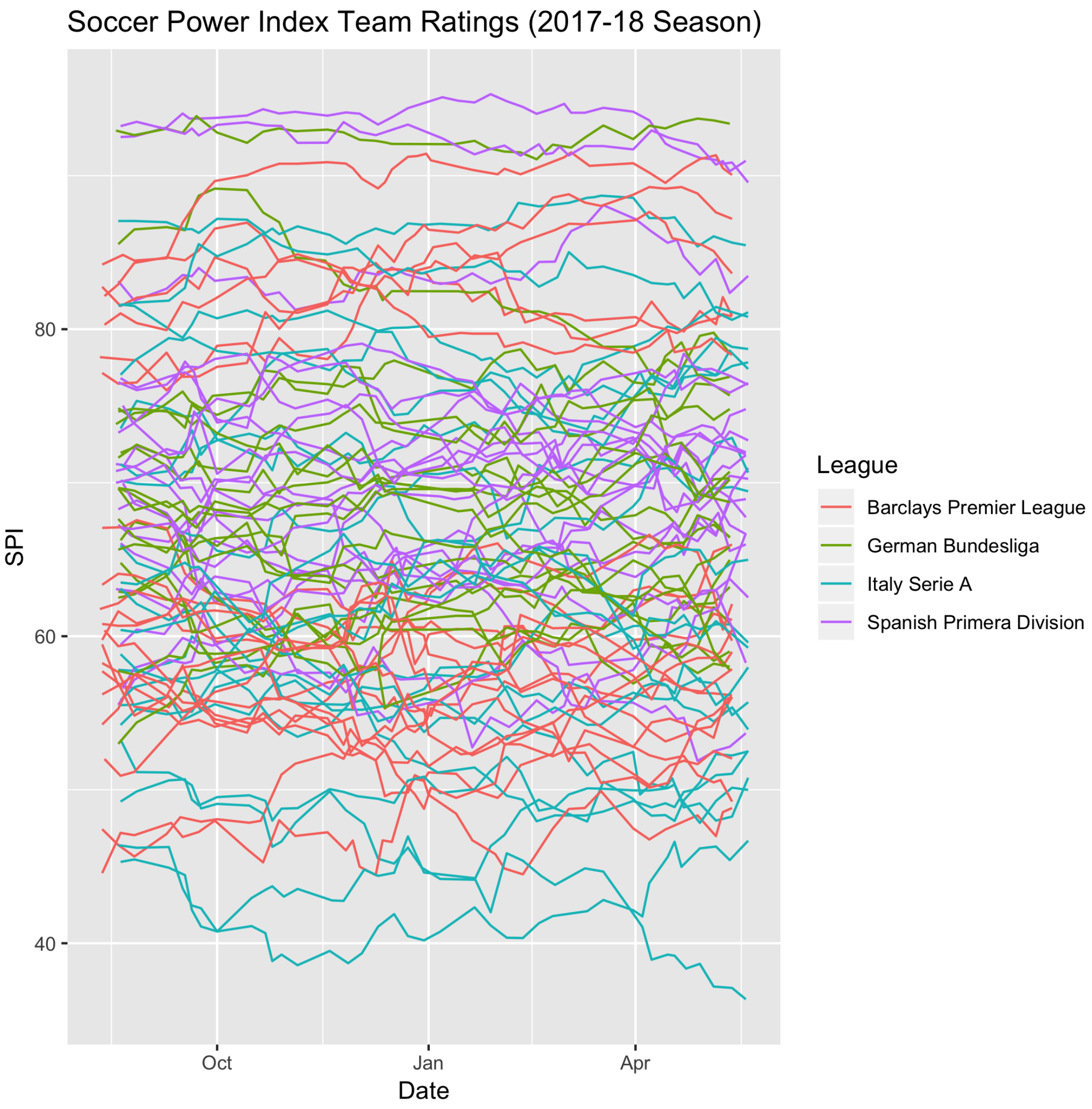

Figure 4: Changes in SPI of clubs over the 2017-18 season.

Appendix B Code

All code and data are available at: https://github.com/aburkard/spi-analysis

~~~
---
title: "SPI Analysis"
output: html_notebook
---

## Load Data

```{r}
spi.table = read.csv("soccer-spi/spi_global_rankings.csv", header = TRUE)
head(spi.table)

trt.deu = spi.table[spi.table$league=="German Bundesliga", "spi"]
trt.esp = spi.table[spi.table$league=="Spanish Primera Division", "spi"]
trt.ita = spi.table[spi.table$league=="Italy Serie A", "spi"]
trt.eng = spi.table[spi.table$league=="Barclays Premier League", "spi"]
```

## BoxPlots

```{r}
library(ggplot2)

spi.table.filtered = spi.table[spi.table$league %in% c("German Bundesliga",
                                                      "Spanish Primera Division",
                                                      "Italy Serie A",
                                                      "Barclays Premier League"),]

spi.table.filtered$league = as.character(spi.table.filtered$league)
spi.table.filtered$league[spi.table.filtered$league == "German Bundesliga"] <-
    "Bundesliga"
spi.table.filtered$league[spi.table.filtered$league == "Spanish Primera Division"] <- "La
    Liga"
spi.table.filtered$league[spi.table.filtered$league == "Italy Serie A"] <- "Serie A"
spi.table.filtered$league[spi.table.filtered$league == "Barclays Premier League"] <-
    "Premier League"
spi.table.filtered$league = as.factor(spi.table.filtered$league)

p <- ggplot(spi.table.filtered, aes(league, spi, fill=league, color=league, alpha=0.4)) +
  geom_boxplot(show.legend = FALSE) +
  labs(title = "Soccer Power Index Team Ratings", x="League", y="SPI")
ggsave("images/boxplot.png")
p
```

## Permutation F-Test

```{r}
source("http://www4.stat.ncsu.edu/~lu/ST505/Rcode/functions-Ch3.R")

set.seed(123)
k = 4
x = c(trt.deu, trt.esp, trt.ita, trt.eng)
n = length(x)
groupLengths = c(
  length(trt.deu),
  length(trt.esp),
  length(trt.ita),
  length(trt.eng)
  )
grps = rep(1:k, times=groupLengths)
~~~

~~~
(trtmeans <- getmeans(x,grps))

#ANOVA based on the assumption of normal distribution with equal variance
summary(aov(x ~ factor(grps)))

# Observed F-value
#Fobs <- summary(aov(x~factor(grps)))[[1]][1,4]
Fobs = getF(x,grps)
perm.F = perm.approx.F(x, grps, R=1000)
(perm.pval = mean(perm.F >= Fobs))
```


## Load across-year data
```{r}
library(dplyr)

spi.matches = read.csv("soccer-spi/spi_matches.csv", header = TRUE)
spi.matches$date = as.Date(spi.matches$date)

spi.matches.filtered = spi.matches[spi.matches$date > "2017-07-01",]
spi.matches.filtered = spi.matches.filtered[spi.matches.filtered$league %in% c("German
    Bundesliga",
                                                  "Spanish Primera Division",
                                                  "Italy Serie A",
                                                  "Barclays Premier League"),]
spi.matches.filtered = spi.matches.filtered[,c("date", "league", "team1", "team2",
    "spi1", "spi2")]
spi.matches.filtered.reverse = spi.matches.filtered
names(spi.matches.filtered.reverse) = c("date", "league", "team2", "team1", "spi2",
    "spi1")
spi.matches.filtered.all = rbind(spi.matches.filtered, spi.matches.filtered.reverse)

spi.matches.filtered.ba = spi.matches.filtered.all %>%
  arrange(date) %>%
  group_by(team1) %>%
  mutate(
    spi.begin = first(spi1),
    spi.end = last(spi1)
    ) %>%
  select(league, team1, spi.begin, spi.end) %>%
  filter(row_number()==1)

spi.matches.filtered.all
spi.matches.filtered.ba
```

## Time series team SPI plot

```{r}
p = ggplot(spi.matches.filtered.all, aes(x=date,y=spi1,colour=league,group=team1)) +
  geom_line() +
  labs(title = "Soccer Power Index Team Ratings (2017-18 Season)", x="Date", y="SPI",
      color="League")
ggsave("images/allTeamsTimeSeries.png",p)
p
```

```{r}
library(gridExtra)
p1 = ggplot(filter(spi.matches.filtered.all, league=="German Bundesliga"),
    aes(x=date,y=spi1,colour=team1,group=team1)) +
  geom_line(show.legend = FALSE) +
~~~

~~~
  labs(title = "Bundesliga", x="Date", y="SPI", color="Club")
p2 = ggplot(filter(spi.matches.filtered.all, league=="Spanish Primera Division"),
    aes(x=date,y=spi1,colour=team1,group=team1)) +
  geom_line(show.legend = FALSE) +
  labs(title = "La Liga", x="Date", y="SPI", color="Club")
p3 = ggplot(filter(spi.matches.filtered.all, league=="Italy Serie A"),
    aes(x=date,y=spi1,colour=team1,group=team1)) +
  geom_line(show.legend = FALSE) +
  labs(title = "Serie A", x="Date", y="SPI", color="Club")
p4 = ggplot(filter(spi.matches.filtered.all, league=="Barclays Premier League"),
    aes(x=date,y=spi1,colour=team1,group=team1)) +
  geom_line(show.legend = FALSE) +
  labs(title = "Premier League", x="Date", y="SPI", color="Club")

grid.arrange(p1,p2,p3,p4, ncol=2)
ggsave("images/timeSeries.png", arrangeGrob(p1,p2,p3,p4))
```

## Paired-Comparison Permutation Tests

```{r}
source("http://www4.stat.ncsu.edu/~lu/ST505/Rcode/functions-Ch4.R")
set.seed(123)

for(league in unique(spi.matches.filtered.ba$league)) {
  spi.matches.leagueFilter <- spi.matches.filtered.ba[spi.matches.filtered.ba$league ==
      league,]
  d <- spi.matches.leagueFilter$spi.end - spi.matches.leagueFilter$spi.begin
  (dbar = mean(d))
  permdbars <- perm.approx.dbar(d, R=10000)
  pval.upper = mean(permdbars >= dbar)
  pval.lower = mean(permdbars <= dbar)
  pval.twotail = mean(abs(permdbars) >= abs(dbar))
  print(paste(league, pval.twotail))
}
```

## MLS / EFL Championship Data

```{r}
trt.mls = spi.table[spi.table$league=="Major League Soccer",]
trt.eng2 = spi.table[spi.table$league=="English League Championship",]

dat <- data.frame(xx = c(runif(100,20,50),runif(100,40,80),runif(100,0,30)),yy =
    rep(letters[1:3],each = 100))
trt.all = rbind(trt.mls, trt.eng2)
p = ggplot(trt.all,aes(x=spi, group=league, fill=league, color=league)) +
  geom_histogram(data=trt.mls, alpha = 0.4, binwidth = 5) +
  geom_histogram(data=trt.eng2, alpha = 0.4, binwidth = 5) +
  labs(title = "MLS/EFL SPI (2017-18)", x="SPI", y="Count")
ggsave("images/mls_efl.png", p)
p
```

## MLS / EFL Championship Test for Deviances
```{r}
source("http://www4.stat.ncsu.edu/~lu/ST505/Rcode/functions-Ch2.R")

x = trt.mls$spi
y = trt.eng2$spi

devx = x - median(x)
devy = y - median(y)
~~~

```
m = length(x)
n = length(y)
rmd2 <- max(mean(abs(devx)) , mean(abs(devy)) )/
      min(mean(abs(devx)) , mean(abs(devy)) )

set.seed(123)
permrmds.approx <- perm.approx.rmd(c(devx,devy),m, R=10000)
mean(permrmds.approx >= rmd2)

rand.perm(x, y, R=1000, alternative = "less", stat= "mediandiff")
rand.perm(x, y, R=1000, alternative = "less", stat= "meandiff")
‘‘‘
```