HW12 - Inference

Stat 131A, Fall 2018

Due Nov-30

Data

In this assignment you will be using the data sets from NBA (seasons 2016-2017 and 2017-2018). The corresponding data files, `nba2017.csv` and `nba2018.csv`, are available in the github repository for this course.

Suggestion: we recommend that you download a copy of the data files to your computer. Here's one way to do that via R (downloading files to your working directory):

```r
# assembling the URL of the CSV files
# (otherwise it won't fit within the margins of this document)
repo = 'https://raw.githubusercontent.com/ucb-introstat/introstat-fall-2018/'
data2017 = 'master/data/nba2017.csv'
data2018 = 'master/data/nba2018.csv'

download.file(url = paste0(repo, data2017))
download.file(url = paste0(repo, data2018))
```

Assuming that the data files are in the directory of your `Rmd` file, you can import the data sets like so:

```r
nba17 <- read.csv('nba2017.csv', stringsAsFactors = FALSE)
nba18 <- read.csv('nba2018.csv', stringsAsFactors = FALSE)
```

1) Proportion of 2-Pointers for Kevin Durant

Consider the season of 2016-2017 (`nba2017.csv`). Say we are interested in studying the proportion of successful 2-point shots of a given player during the entire length of time of the regular season (including games as well as practice and training sessions).

The issue is that the available data contains only shots attempted-and-made during the 82 games of the regular season. But we don't have data for the attempted-and-made shots during the practice, training, and warm-up sessions.

For example, the number of 2-point shots attempted by LeBron James, during the games, was:

```r
lebron <- which(nba17$player == 'LeBron James')
nba17$points2_atts[lebron]
```

```
## [1] 1002
```

Likewise, the number of 2-point shots (successfully) made by James was:

```r
nba17$points2[lebron]
```

```
## [1] 612
```

This gives us a proportion of successful 2-point shots of:

```r
nba17$points2[lebron] / nba17$points2_atts[lebron]
```

```
## [1] 0.6107784
```

If we assume that the attempted 2-point shots during the games of the regular season represents a random sample of ALL the *unobserved* 2-point shots (shots during both the games, and outside the games), we can consider the above proportion as the *sample proportion* for LeBron James, denoted as $\hat{p}_{james} = 0.6107784$.

Your turn:

a) Estimate 90%, 95%, and 99% confidence intervals for the proportion of all successful 2-point shots of Kevin Durant.

b) How large a sample is needed to ensure that the 95%-confidence interval estimate of p_{durant} is less than 0.05? (i.e. margin of error less than 0.05)

2) Proportion of 2-Pointers for all players

In total (i.e. between all players), there were 128459 2-point shots attempted during the 82 games in the regular season 2016-2017. Assume that these shots represent a random sample of ALL the *unobserved* 2-point shots attempted by all NBA players during the entire regular season (not only games but also practice/training/warm-up sessions).

a) Estimate 90%, 95%, and 99% confidence intervals for the proportion of successful 2-point shots for all NBA players during the 2016-2017 season.

3) Stephen Curry's 3 Pointers

Suppose that during the 2016-2017 regular season, the proportion of ALL (unobserved) successful 3-point shots of Stephen Curry (point guard for the Golden State Warriors) was 0.40.

a) What was the (observed) proportion of successful 3-point shots scored by Stephen Curry during the games in season 2017-2018?

b) A basketball analyst claims that the percentage of successful 3-point shots made by Stephen Curry during 2017-2018 "significantly" increased from the previous season 2016-2017.

Assume that the 3-point shots attempted-and-made by Stephen Curry during the games of 2017-2018 represents a random sample of ALL his unobserved 3-point shots during that season. If we consider a significance level of 5%, does the data from 2017-2018 support the claim made by the analyst?

Perform a hypothesis; state the null and alternative hypotheses; determine and compute the test statistic; calculate the p-value; and make a conclusion.

4) One-sample *t*-test

We are interested in understanding annual salaries of basketball players who play for the Golden State Warriors (`nba17$team == GSW`), so we investigate the variable `salary` in our `nba2017.csv` dataset.

a) Suppose we decide to conduct either a 1-sample z-test or a 1-sample t-test. What assumptions does each test require?

b) Get graphical and numerical summaries of the salary data for the Golden State Warriors in 2017. In particular, comment on whether to conduct a t-test or a z-test.

c) Regardless of your answers above, compute the z-statistic to the test the null hypothesis: the Golden State Warriors annual salary is greater than 1 million. Carry out the test of significance and provide an interpretation of the resulting test statistic. Be sure to clearly specify the how the test statistic was computed, and what the rejection region is.

d) Now, compute the t-statistic for the same null hypothesis. Carry out the test of significance and provide an interpretation of the resulting test statistic. Be sure to clearly specify the how the test statistic was computed, and what the rejection region is. Assuming the null hypothesis is false, **which test provides greater power**?

5) Paired t-test and two-sample t-test

Many NBA basketball players played in both the 2017 and 2018 seasons. Their corresponding salaries in each season can be obtained as:

```r
# salaries of those who played in both seasons (2017, 2018)
# (values already matched by player name)
salary17 <- nba17$salary[which(nba17$player %in% nba18$player)]
salary18 <- nba18$salary[na.omit(match(nba17$player, nba18$player))]
```

a) For each player who played in both years, find the difference in salary. Provide numerical and graphical summaries. Comment on whether a test of significance on year-to-year salary difference should be one- or two-sided.

b) Conduct a test of significance based on the t-distribution to test the null hypothesis: the expected change in salary for a basketball player in the NBA is 0. Be sure to clearly specify what test of significance should be conducted, how the test statistic was computed, and what the rejection region is.

c) Now suppose that an NBA player's salary one year is independent of their salary in any other year. Conduct a test of significance based on the t-distribution to test the null hypothesis that the expected annual salary of an NBA player is constant year to year. **Comment** on the extent to which the choice of test of significance results in different results, and **provide a brief explanation**. Be sure to clearly specify what test of significance should be conducted, how the test statistic was computed, and what the rejection region is. Assume that the unobserved population distributions of NBA salaries each year have **equal variance**.