Hypothesis Testing

In this first example we're going to accomplish 2 things:

- See hypothesis testing in action
- Use R to make the calculations

The example is intentionally easy, so that we can concentrate on the technique, more than on understanding the business logic being it. In another section of this website, I'll try to concentrate myself more on business relevant questions and hypothesis.

As we'll always do, we start from the formulation of a relevant question. In this particular case we'll be dealing with the MLB (Major League Baseball). We'll pretend to be interested in formally answering the question:

> *are the teams making the postseason spending more money for players salaries than those who don't?*

I know, I know, it's a pretty rethorical question, but indeed, this is intentional. To answer it we'll use the History of Baseball data sets provided by Kaggle: it contains a complete history of major league baseball stats from 1871 to 2015, just what we need to have some fun with data!

The method

First of all we state the test hypothesis and set parameters.

Under the **NULL Hypothesis** (*H0*) we state that total amount payed in salaries by the teams who qualify for the playoffs, is the same as those payed by teams who don't qualify for the playoffs, whereas under the **Alternative Hypothesis** (*Ha*) we state that the teams who qualify for the playoffs pay higher salaries. In order to set up these hypothesis we need to divide the population of all MLB teams from 1871 to 2015 into two groups: postseason_teams and regular_season_teams and calculate the average of total salaries for the two groups. These are the quantitative variables that will be used for our hypothesis

- $\mathrm{H}_0 : \mu_{ps} = \mu_{rs}$
- $\mathrm{H}_a : \mu_{ps} > \mu_{rs}$

Note: ps = Post Season Teams, rs = Regular Season Teams (didn't make it to the playoffs)

This is going to be right-sided test, because of the Alternative Hypothesis.

We set a **Level of Significance** *alpha* for the test of 0.05.

A comment to the method

In hypothesis testing we start by stating that the NULL Hypothesis is true.

Our goal though is to supply enough evidence in order to demonstrate that it isn't.

The logic behind that is the following: using the sample data that we have at disposition we calculate, by means of the test statistic that we'll see in the next section, how likely it is that we would draw the observed samples from a population if the NULL Hypothesis were true.

If this is very unlikely, than our drawing these two very different samples must not be due to chance or to sampling error, but to a real effect present in the population, so we reject the NULL Hypothesis of equality in favor of the Alternative Hypothesis of inequality of the two groups.

The Data Preparation

We'll use 3 data sets:

- team.csv: data about all teams for each year
- postseason.csv: data about all postseason teams for each year
- salary.csv: data about salaries for each player, from 1985 to 2015

The preparation steps are essentially three:

1. Calculate adjusted salaries taking the CPI (consumer price index) into account. All salaries will be brought to their equivalent 2015 level
2. Select only the necessary variables from the source files:
    - year
    - team_id
    - name: team name
    - PCT: percentage of victories
    - salary: original salary cap
    - cpi: consumer price index
    - adj_salary: adjusted salary cap to the 2015 consumer price index
3. Divide the teams into two groups: those who qualified for the postseason and those who didn't.

The final data sets should look something like this:

```
head(all_teams, 5)
```

```
##   year team_id                name   PCT   salary      cpi adj_salary
## 1 1985     KCA  Kansas City Royals 0.562  9321179 165.6917       19.6
## 2 1985     LAN Los Angeles Dodgers 0.586 10967917 165.6917       23.0
## 3 1985     SLN St. Louis Cardinals 0.623 11817083 165.6917       24.8
## 4 1985     TOR   Toronto Blue Jays 0.615  8812550 165.6917       18.5
## 5 1986     BOS      Boston Red Sox 0.590 14402239 168.6500       29.7
```

The Test

Once the data sets are ready we start applying the test.

First of all we need to check and see, whether the data have the necessary preconditions to run the test.

The **first condition** is that the **samples are independent** and we will consider them so, even if this may be a bit far fetched.

The **second condition** is that **sampling distributions are approximately normal**. We can check that directly, letting R draw the histograms of the two groups.

They look normal enogh, for the purposes of this demonstration, even if the Regular Season Salary distribution has a right skew.

In order to give NULL and Alternative Hypothesis real numbers, we need the mean salaries of each group, and most importantly their difference. The statistical test is a comparison of the means between these two groups and an attempt to see if there is enough statistical evidence to conclude that the observed difference is due to an actual effect in the population (of all possibile infinite MLB teams), or if it is due to sampling error and chance.

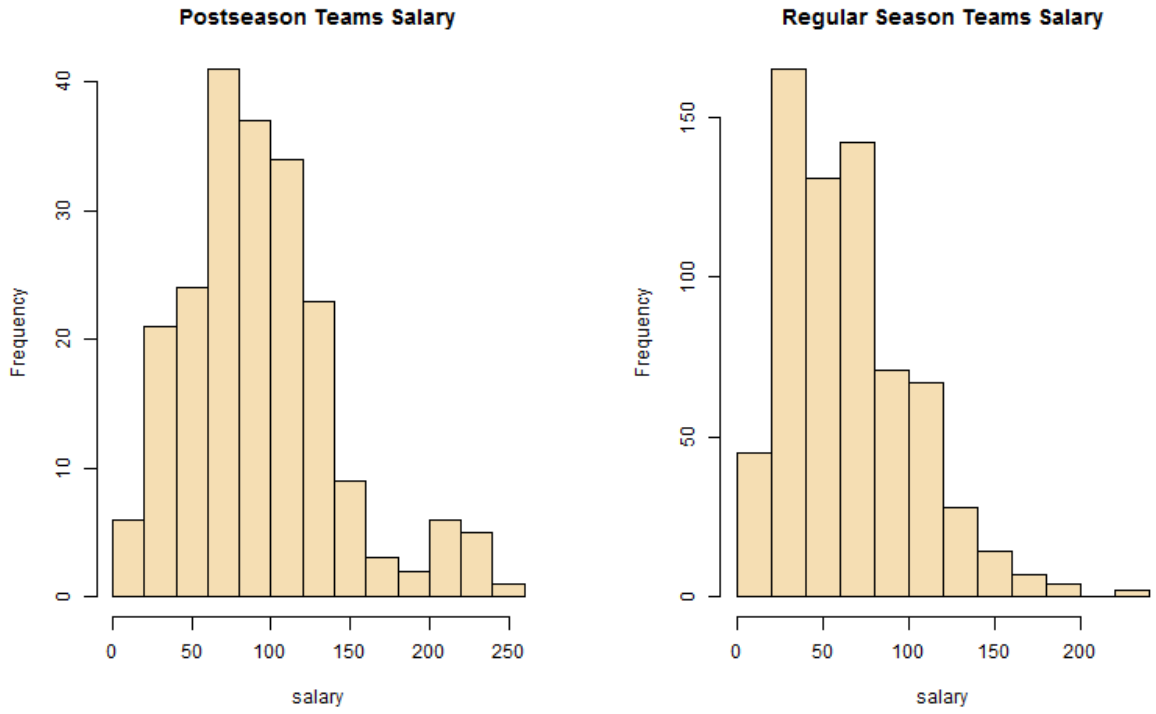

Figure 1:

```
mu_ps <- mean(postseason_teams$adj_salary)
mu_rs <- mean(regular_season_teams$adj_salary)
difference <- mu_ps - mu_rs
three_stats <- round(c("Avg. Sal. Postseas" = mu_ps, "Avg. Sal. Regseas" = mu_rs, "Difference" = differe
three_stats
```

```
## Avg. Sal. Postseas  Avg. Sal. Regseas         Difference
##             93.065             64.559             28.505
```

The Hypothesis, reformulated are

- $\mathrm{H}_0 : difference = 0$
- $\mathrm{H}_a : difference > 0$

We perform a t-test.

The formula for the test statistic is

Here's the evaluation of all required variables

```
## [1] "Sample size of postseason teams = 212"
```

```
## [1] "Sample size of regular season teams = 676"
```

```
## [1] "Standard Deviation of postseason teams salary =  48.285"
```

$$t = \frac{difference}{\sqrt{\frac{sd_{ps}^2}{n_{ps}} + \frac{sd_{rs}^2}{n_{rs}}}} = \frac{difference}{standarderror}$$

Figure 2: test-statistic

```
## [1] "Standard Deviation of regular season teams salary = 37.193"
```

```
## [1] "Standard Error (Denominator of the formula) = 3.612"
```

Which leads to a t score of

```
## [1] 7.89
```

Now there are 3 ways to use this test statistics in order to either reject the NULL Hypothesis in favor of the Alternative Hypothesis, or fail to reject it and thus fail to reach evidence:

1. we calculate a p-value and compare it to the given Level of Significance. If the p-value is less than the Level of Significance, then we reject H0, oherwise we fail to reject it.
2. We calculate the cut-off value and see whether our test statistic falls within the rejection region. If it does, then we reject H0, otherwise we fail to reject it.
3. We calculate a confidence interval for the difference of the means and see whether it contains the value 0. If it doesn't, then we reject H0, otherwise we fail to reject it.

They all lead to the same conclusion and we'll test them all.

The **p-value**:

```
## [1] "With a p-value of 5.59021465468716e-14 and a Confidence Level of 0.05 : Reject HO"
```

The **cut-off value**:

```
## [1] "With a cut-off value of 1.651 a right-sided test and a t-value of 7.89 : Reject HO"
```

We can see graphically too, that the t-value falls in the rejection region

The **confidence interval**:

```
## [1] "With a 95% Confidence Interval of [ 21.39 , 35.62 ]: Reject HO"
```

We can see graphically, that 0 is indeed not contained in the Confidernce Interval, thus allowing us to rejectH0.

As was expected, by rejecting the NULL Hypothesis we reach evidence that, indeed, the teams who qualify for the playoffs spend more money on their salary caps.

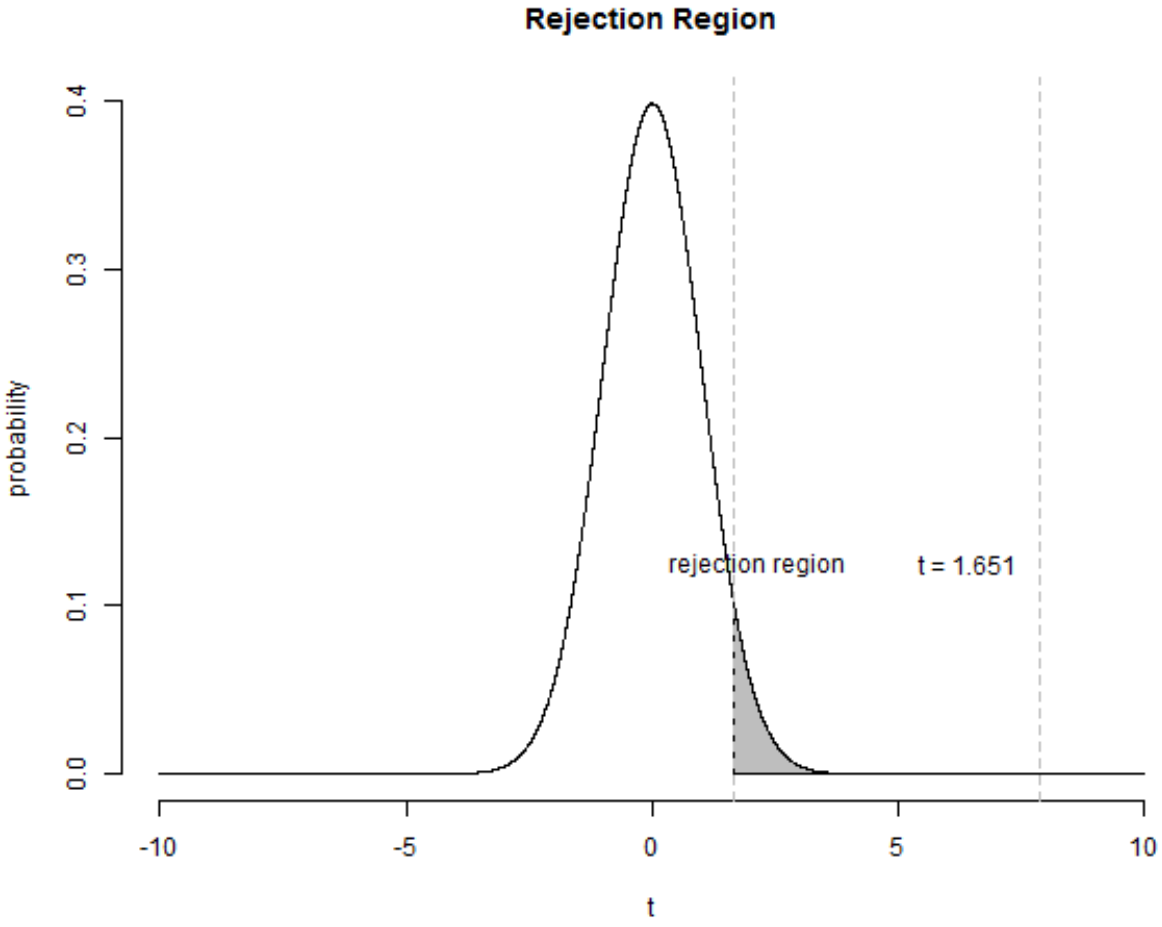
Rejection Region
probability
0.0
0.1
0.2
0.3
0.4
rejection region
t = 1.651
-10
-5
0
5
10
t

Figure 3:

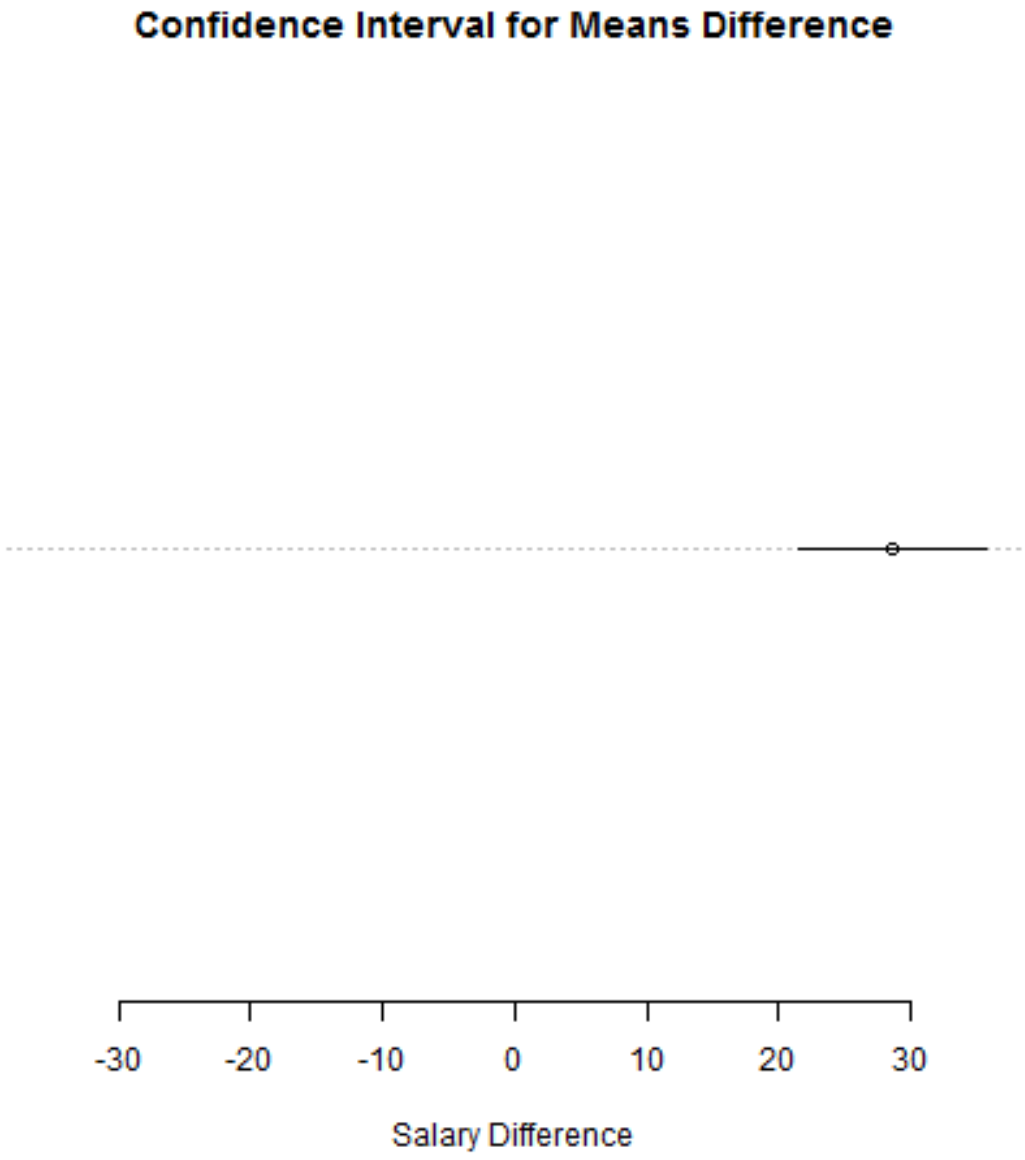
Confidence Interval for Means Difference
-30
-20
-10
0
10
20
30
Salary Difference

Figure 4: