# 5. Hypotheses, Success Definition, and Methodology

The goal needed to be defined by which could be investigated and analyze the research questions using the data. The fact that the concept of "success" was interpreted differently in each situation, for research purpose the success index was divided into 4 categories, each category named as label:

1. Amazing Success:

No revolving door cases, placement since enters the program = 1, no resumption date.

1. Medium Success:  
   No revolving door cases, placement since enters the program > 1, no renew activity date, or no renew registration date.
2. Weak Success:

All job seekers who not in label 1,2 or 4.

1. Failure:

No resumption date, no placements since joining the program, or revolving door cases higher than 0.

Research questions:

(Q1) Is there a difference between good placements of the Arabic population to other populations?

(Q2) How long job seekers are in the program before their placements?

(Q3) Is there a difference between the time length of jobseeker in the program to type of placement?

(Q4) Do the number of activities from the program effects job seeker's placements?

(Q5) Is there an effect between socio-economic jobseeker characteristics to placement?

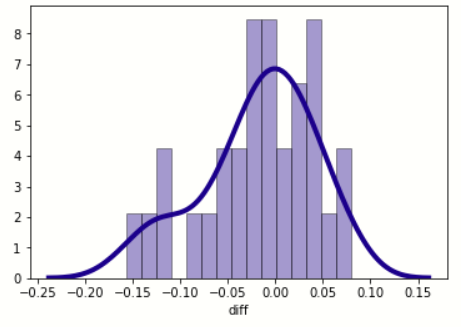
## *Q1: Is there a difference between good placements of the Arabic population to other populations?*

In 2015, the percent of labor market employment in the Arabs community was 54.6% and 81.7% in the Jews community ([משרד העבודה והרווחה, 2015](#משרד_העבודה_והרווחה_2015)) and in 2016 the Arabs employment rate was 42.5% ([הלשכה המרכזית לסטטיסטיקה, 2017](#שנתון_סטטיסטי_לישראל_2017)) The number of unemployed Arabs is larger and requires action. Israeli government makes affirmative actions to Arabic people and gives funds for employment places to grow, a fund for education (scholarships, reduced taxes, etc), and more. There are many reasons for affirmative actions like cultural differences, traditions, geographical environment, social status, and more In the IES they have government funds (1.2 million NIS) for a specific program called "Tapuah" (the program gives 1.466 million NIS) ([שירות התעסוקה ועמותת תפוח, 2017](#תפוח_2017)) specific organized for Arab people, but there is no evidence that will justify open special programs to their community at this program expense, or at least show that 'Employment Circuits' has a negative or non-effect for Arabs. The hypothesis that there is no difference in their placement in the 'Employment Circuits' program will be examined with .

(1)

There is a need to see if the data is distributed of normal distribution. Therefore, we need at least 30 bureaus where the number of Arabs will be a statistical basis for the hypothesis (at least 14 Arab jobseekers in the bureau needed to reach 30 unique bureaus). A comparison was made between the Arabs with good placement and the non-Arabs with a good placement from the same bureau. The 'diff' column was calculated, and it's representing the differences in good placing proportion to estimate the differences in placements in the same bureau. There is a need to check if the differences are normally distributed, the histogram will show the number of samples relative to the difference and by a density function.

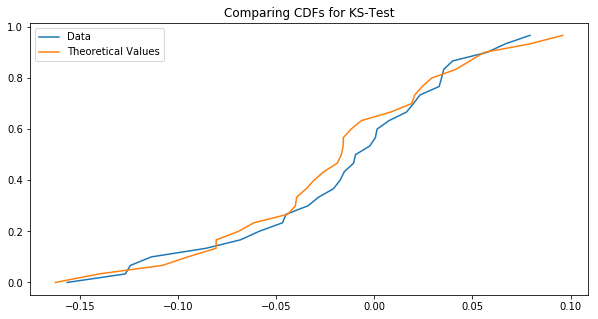
*Figure 1. Histogram of 'diff' column*



(The X-axis shows the differences and the Y-axis shows the number of samples.  
The estimate of variance is 0.0035).

It seems from the data that the density function (blue line) does imply that the data is a normal distribution (obviously this is not enough, and a statistical test will be performed below).

*Figure 2. Comparing CDFs for KS-Test*



After comparing CDFs for KS-Test, it seems there is a very large resemblance to a normal distribution, so to verify that the good placement differences are distributed normally Kolmogorov Smirnov test was made via SPSS to confirm.

***Table 3****Test of Normality*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
| Statistic | df | Sig. | Statistic | df | Sig. |
| diff\_placments | .116 | 30 | .200\* | .956 | 30 | .244 |
| \*. This is a lower bound of the true significance. | | | | | | |
| a. Lilliefors Significance Correction | | | | | | |

Kolmogorov Smirnov and Shapiro-Wilk test results show ρ (the critical value) is greater than ∝ so we say the diff column is normally distributed. since we have 30 samples and we saw that the data is normally distributed we can use a t-test to examine the difference between dependent pairs (30 bureaus out of 60 bureaus). It is assumed that there are more similar characteristics to populations from the same bureaus.

The expected value estimator is: and the result is -0.018.  
Calculation of our statistic: and its results are -1.686.  
Therefore, T-test result is: t (0.95,29) = 1.699, ,.

The conclusion is to reject the null hypothesis and say there is no significant difference between the placement of the Arabs and the placement of the non-Arabs.

## *Q2: How long job seekers are in the program before their placements?*

To answer this question, a copy of the data was made into a new file and transfer to the data frame with the columns of "Last Placement Date", "Last Placement Report Date", " Initial Entry Date". Some data have blank dates ("NAT"). To understand how long the jobseekers in the program, a calculation was made in the 'day\_diff' column.

***Table 3*** *'Days\_Diff' division to categories*

|  |  |  |
| --- | --- | --- |
| Percent of job seekers | Number of job seekers | Category |
| 0.366% | 205 | 0 days |
| 2.865% | 1604 | 1 – 10 days |
| 5.067% | 2842 | 11 – 30 days |
| 5.564% | 3115 | 31 – 60 days |
| 4.574% | 2561 | 61 – 100 days |
| 4.081% | 2285 | 101 – 150 days |
| 3.074% | 1721 | 151 – 200 days |
| 7.212% | 4038 | 201 – 365 days |
| 17.191% | 9625 | Above a year |
| 49.997% | 27993 | None |

Almost half (49.99%) of program participants do not have a "placement date" and more than 25% have "first entry date" greater than "last placement date".

After clearing the records of the negative diff, the data were summarized. Confidence interval is: (Left: [336.994], Mean: 341.418, Right: [345.8417])

## *Q3: Is there a difference between the time length of jobseeker in the program to type of placement?*

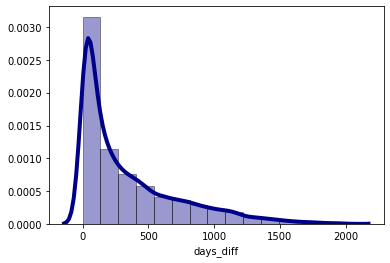
A continued question to the previous result, is there a difference between the time length of jobseeker in the program to type of there placements? Is longer attended to this program give more tools to jobseeker to find a job, but is this the reality? And if it is, a longer stay needs to return a better placement. Another hypothesis is there is a difference between the time of job seekers in the program to the type of their placements, the examination will be examined with  Figure 6.3.1 shows the division of different labels and the length of their being in the program.

*Figure 3. Time of jobseeker in the program division by label*



[Figure 3](#Figure_3) hint's that better label has a shorter time frame in the program (vice versa to the thought of longer stay need to return a better placement). But it is not enough just in the graphs of the data to conclude this, a statistical test must be performed to verify it.

*Figure 4. 'days\_diff' column distribution*



To test if the data is normally distributed, the Kolmogorov Smirnov test was made and the result of ρ is ~0.0, therefore, we do not reject the null hypothesis.

***Table 4*** *Kolmogorov-Smirnov test result*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Kolmogorov-Smirnova | | |
| Statistic | df | Sig. |
| days\_diff | .183 | 27996 | .000 |
| a. Lilliefors Significance Correction | | | |

To check the difference between the day's differences between the different labels an ANOVA test was done, and its results are within label there is a difference between the different label groups. To know which label has a difference with another, a Tukey post hoc test was made to test time differences between each label pair separately.

***Table 5*** *Tukey Post Hoc Test*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| hedges | ρ | T | tail | se | diff | Mean(B) | Mean(A) | B | A |
| -0.286 | ~ 0 | -11.964 | two-tail | 8.574 | -102.577 | 229.087 | 126.509 | 2 | 1 |
| -0.479 | ~ 0 | -19.723 | two-tail | 8.705 | -171.698 | 298.207 | 126.509 | 3 | 1 |
| -0.947 | ~ 0 | -41.959 | two-tail | 8.087 | -339.326 | 465.836 | 126.509 | 4 | 1 |
| -0.193 | ~ 0 | -11.061 | two-tail | 6.249 | -69.121 | 298.207 | 229.087 | 3 | 2 |
| -0.661 | ~ 0 | -44.216 | two-tail | 5.354 | -236.749 | 465.836 | 229.087 | 4 | 2 |
| -0.468 | ~ 0 | -30.133 | two-tail | 5.563 | -167.628 | 465.836 | 298.207 | 4 | 3 |

Clearly, we can say confidently that between all the means pairs exist differences such that the means declares with the decrease of the labels, and this is based on the two test results that were hypothesized.

## *Q4:* *Do the number of activities from the program effects job seeker's placements?*

Only 30 unique activities defined in the data frame. It was necessary to change the "Activities in the program" column to a categorical variable and layout the activity in a new column and to provide a binary classification if the job seeker was in the activity = 1 if not 0.

Checking the hypothesis that more activities job seekers will anticipate cause decreasing labels for job seekers.

*Figure 5. Number of programs per label*

After these actions, the sum of all activities in which the label was assigned would be summed and see the percentage of all activities.

***Table 5*** *Sum and percent of jobseekers from each label and activity*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Label | Change course | Process course | Occupational Hebrew course | Computer application course | Personal training | … | Number of programs |
| 1 | 131 | 3 | 42 | 0 | 1231 |  | 2457 |
| 2 | 507 | 10 | 133 | 1 | 4200 |  | 8643 |
| 3 | 395 | 7 | 72 | 0 | 3786 |  | 7435 |
| 4 | 2550 | 60 | 674 | 14 | 23151 |  | 48900 |
| Label 1 Fraction | 4% | 4% | 5% | 0 | 4% |  | 4% |
| Label 2 Fraction | 14% | 12% | 14% | 7% | 13% |  | 13% |
| Label 3 Fraction | 11% | 9% | 8% | 0% | 12% |  | 11% |
| Label 4 Fraction | 71% | 75% | 73% | 93% | 72% |  | 73% |

According to the table above, anyone with label 1 seems to greatly reduce their number of activities in general and in each activity from any other label. It was necessary to see if there was a significant difference in the variance between activities in the program with the ANOVA test, but first to check if the data is normally distributed. Kolmogorov Smirnov test showed significant results with each number of programs per label, therefore, ANOVA test was performed and the of ANOVA is ~0 which means its significantly shown that it is different between different activities and the time when job seekers are on the program. It was necessary to continue investigating the difference between activities using the Tukey Post Hoc test.

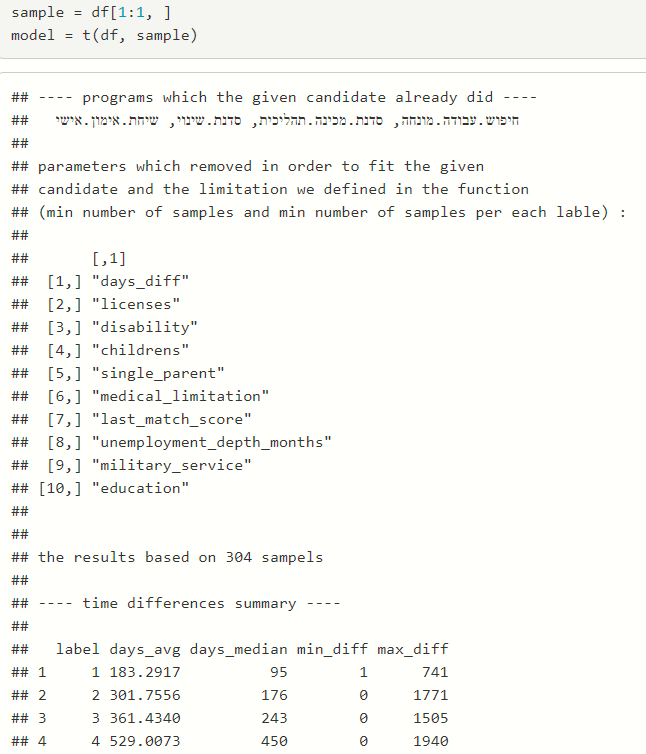
***Table 5*** *Tukey Post Hoc test*

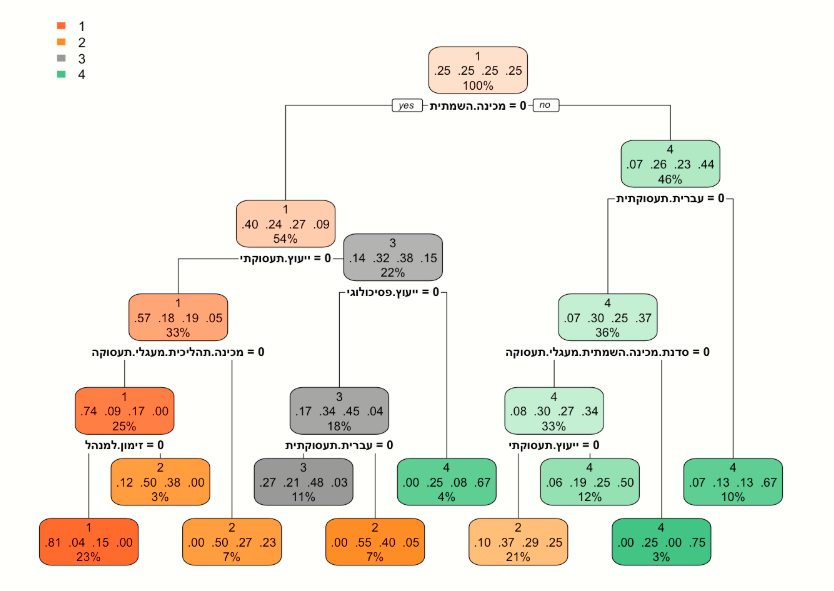
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| hedges | ρ | T | tail | se | diff | Mean(B) | Mean(A) | B | A |
| -0.12 | 0.001 | -5.251 | two-tail | 0.033 | -0.174 | 1.14 | 0.967 | 2 | 1 |
| -0.082 | 0.002 | -3.548 | two-tail | 0.034 | -0.119 | 1.068 | 0.967 | 3 | 1 |
| -0.198 | 0.001 | -9.696 | two-tail | 0.03 | -0.287 | 1.253 | 0.967 | 4 | 1 |
| 0.038 | 0.103 | 2.277 | two-tail | 0.024 | -0.055 | 1.068 | 1.14 | 3 | 2 |
| -0.078 | 0.001 | -6.224 | two-tail | 0.018 | -0.113 | 1.253 | 1.14 | 4 | 2 |
| -0.116 | 0.001 | -8.862 | two-tail | 0.019 | -0.168 | 1.253 | 1.068 | 4 | 3 |

There is a difference in the number of activities between the different types of successes (label), except for groups 2 and 3.

After exploring all the different activities as individuals, a decision tree model that would take a sample of jobseekers and match their socioeconomic characteristics and activities against the data frame from which it was sampled will be helpful to see it graphicly. The model models jobseekers and presents the most definite trajectory for him according to the jobseekers he resembles, both in terms of their programs and in terms of socioeconomic characteristics. The decision to use the decision tree is because it shows the most recommended route (order of action). There are many other models for this kind of decision, but after considering alternatives it was decided that this model would be best for the given situation. The model shows the most definite trajectory, so the tree should be considered as the order of best practice for those sampled job seekers. Of course, the tree is no definite promise that according to the proposed route, job seekers will be implemented in this way.

*Figure 6. Decision tree sample result*





## *Q5: Is there an effect between socio-economic jobseeker characteristics to placement?*

This question was asked to see if there is a socioeconomic characteristic that affects the placement more than other characteristics. The definition of socioeconomic characteristics of jobseekers is religion, age, single parent, gender, level of education, city, language, country of birth, marital status, children up to age 18, classification of the jobseeker, disability rates, medical disability, licenses, military service, released prisoner and month of placement ("last placement date"). After building a new data frame that consists of these columns, a sorted data was insert. Uniting religions into major religions: Jewish, Christian, Muslim, and Druze, otherwise, it will have value: another. Level of education was categorized into major: elementary, high school, degree, and professional certificate. Ages were grouped according to the employment service practice: 15-29, 30-39, 40-49, 50-54, 55+. In the 'Land of Birth' column, there are several places in Israel that were divided: Israel, Judea, and Samaria, Golan Heights. Any country of birth with less than 1% frequency dropped from the data because they are end-cases that will not affect the results below. Using a multinomial regression model comparing the model with all variables to the cutter model only.  
McFadden's formula, subtracting from the estimate log distribution from 1, therefore, the higher the resulting value (between 0 and 1) the more pronounced the model is.

It is important to understand that multinomial multivariate regression makes it difficult to reach a result close to 1 because very strong explanatory parameters are needed to increase the value of McFadden's estimate. In attempting to play with the model and to interplay variables and remove irrelevant variables, no more statistically significant result was obtained than the full model. In the case of my data, the McFadden estimate is 0.239. According to McFadden the estimate for a good fit model is between 0.2 and 0.4 [12], so it can be said to be statistically significant and that there is a correlation between the socioeconomic variables and the type of placement. The variables presented by the model are more likely to influence the placement result (which type of label is jobseeker).

Some of the model results can be present as commonsense thinking, but some can surprise us all. All model results are compared to label 4 (failure) and each model result was displayed is significant (ρ < 0.05).

1. There are more Arab cities compered to Jewish cities in labels 1 and 2.  
   Muslim and Jew religions significantly more to be in label 1.  
   All that supporting question 1 results and conclusion.
2. Some country origin effects on program success, like the Soviet Union and France significantly more to be in label 3 than label 1 (compered by p-value).  
   Ethiopia significantly more to be in label 2.
3. Education has a strong effect on the success label. Academic degree or 'Teudat Bagrut' has significantly more to be in label 1, and non-education significantly more to be in label 3.
4. Disability affects human life including program success. From 20%-59% significantly more to be in label 2 and 60%-100% to be in label 3, but no disability at all significantly more to be in label 2 and not 1 as we expected.