Inferential Statistics Report for Building Data

Introduction

The main objective of the capstone project "Predicting Permit Issuance times in San Francisco", is in its title. However, the data gives some more interesting insights which can be verified with a few well-known statistical methods.

Data: Data for this experiment is cleaned version of the building data downloaded from open data portal of city of San Francisco. More details about that is available in Data Wrangling Report. The data contains permit applications with their dates, types, location, construction types and many more info for the city of San Francisco for the duration Jan 2013-Sept 2017. The data has more than 150,000 records.

We keep in mind that data we have is the entire population of year 2013 onwards. Hence the tests have to be performed on a randomly chosen subset assuming entire population is not available. We chose a random sample of size 7500 without replacement and conducted all the tests.

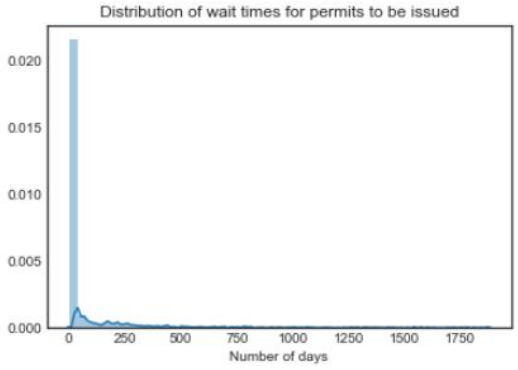
This Document

This document provides results of a few Inferential statistics tests conducted on a **sample** taken from the data.

1) Is it true that on an average 65% of the applications are processed the same day?
Assume there is a claim by the DBI. "On an average 65% of **all** the permit applications are processed on the same day." Can we formulate a hypothesis test? What are the null and alternate hypothesis? What is the significance level you want to chose?

Answer: This hypothesis test is of the type, 1 sample t-test for population proportion. Before conducting the test, we verify certain assumptions for tests for proportions:

- Independence: The samples are independent and without replacement, with sample size < 5% of population
- Sample size/skew: Expected value of success: np >= 10, Expected value of failure: n(1-p)>=10 are satisfied for n=7500, p=0.6
- Distributions nearly normal? No, but this is not important as n is much larger than 30.



Population distribution, which is assumed to be represented by sample distribution as shown above is not nearly normal. However, since the size is much larger than 30, the central limit theorem kicks in. Below we formulate the hypotheses

Null Hypothesis Ho: DBI **processes 65% permit** applications the same day.

Alternate Hypothesis HA: DBI normally processes **a different percentage** of applications **than 60%** the same day and difference is statistically significant.

Note that DBI can not process less than 0% applications.

This is a two sided t-test which is same as z-test due to large size.

Let us have significance level **alpha = 0.01**

```
# The Hypothesis test code. time_taken_sample is the array containing wait times.
# Claimed by DBI, the null proportion
claimed_proportion = 0.65
# Observed proportion, the alternate proportion
observed_proportion = (time_taken_sample == 0).sum()/time_taken_sample.shape[0]
print("Observed proportion of permits issued same day is {:.2f}% ".format(100*observed_proportion))
```

```
Observed proportion of permits issued same day is 59.59%
```

```
from scipy import stats
zsc = (claimed_proportion - observed_proportion)/np.sqrt(claimed_proportion*(1-claimed_proportion)/n)
print("Z score: {:.2f}".format(zsc))
pval_test1 = 2*(1 - stats.norm.cdf(abs(zsc)))
print("p-value is",pval_test1)
```

```
Z score: 9.83
p-value is 0.0
```

Looking at low p-value, lower than alpha, we reject null hypothesis in favor of alternate hypothesis.

This means,

“The probability of observing a percentage value for permits processed, as observed with this sample, given null hypothesis is true is less than 3e-05”, hence there’s enough evidence to believe that null hypothesis is false.

1a) Calculate the 95% confidence interval for percentage of permits that get processed the same day.

Answer: We calculate the margin of error corresponding to 95% confidence interval and then add to and subtract from the observed proportion, to get a range for population proportion. This range is called confidence interval. It means, if 100 times a sample of size 7500 is drawn from the population, it is likely that 95 of such samples will yield a percentage permits issued that lies in the range mentioned by 95% confidence interval.

```
# Z score for 95% CI is 1.96
z_95 = 1.96
# From the formula for margin of error,
margin_error = z_95 * np.sqrt(observed_proportion*(1-observed_proportion)/n)
print("Margin of error for population proportion: {:.4f}".format(margin_error))
# By definition, confidence interval:
ci = (100*round(observed_proportion-margin_error,4),100*round(observed_proportion+margin_error,4))
print('95% confidence interval of the population proportion:',ci,'%')
```

```
Margin of error for population proportion: 0.0111
95% confidence interval of the population proportion: (58.48, 60.699999999999996) %
```

Refer this * for page for short explanations and formulae.

2) Is there a statistically significant difference between mean wait times of fire only permits and not fire only permits?

Answer: This hypothesis test is of the type, 2 sample t-test for population means.

Ho: Average wait times are same whether it is fire only permit or not

HA: Average wait times are different

Again alpha = 0.01

Sample statistics:

fire_only_permit	count	mean	std	min	25%	50%	75%	max
N	6800.0	72.542647	227.740143	0.0	0.0	0.0	13.0	1847.0
Y	700.0	32.514286	174.737804	0.0	0.0	0.0	2.0	1829.0

Before conducting the test, we verify certain assumptions for 2 sample tests:

- Independence: These two samples are independent within themselves and of each other
- Sample size: Both of them have large enough (>30) count and they are less than 5% of the population, taken without replacement.
- Distribution: They are not nearly normal, but that is fine due to large size.

```
# Compute combined standard deviation
comb_sigma = np.sqrt((sample_stats.loc['N','std']**2 / sample_stats.loc['N','count'])
                     + (sample_stats.loc['Y','std']**2 / sample_stats.loc['Y','count']))

# Z-score
zsc2 = (sample_stats.loc['N','mean'] - sample_stats.loc['Y','mean'])/comb_sigma
print("Z score for 2 sample z test: {:.2f}".format(zsc2))

# p-value
pval_test2 = 2*(1 - stats.norm.cdf(abs(zsc2)))
print("p-value for 2 sample test is",pval_test2)
```

```
Z score for 2 sample z test: 5.59
p-value for 2 sample test is 2.2497833596091255e-08
```

Here due to close to zero p-value, we reject null hypothesis in favor of alternate hypothesis.

3) Can we do ANOVA test to verify variability across mean wait times for various permit types? If yes, demonstrate. If no, justify.

Answer: ANOVA test is done to see if the difference between statistic of more than two populations is statistically significant or not. That is,

Ho: There is no difference between statistic of N different populations.

HA: At least one of the populations differs from the remaining.

With alpha = 0.01

The conditions for doing ANOVA test:

- The observations should be independent within and across groups : This is met.
- The data within each group are nearly normal: This is not met as can be seen from the asymmetric percentile values in the above table.

- The variability across the groups is about equal (and use graphical diagnostics to check if these conditions are met.): Without graphical diagnostics, it is evident from the "std" column that
- variability is not met.

Conditions required for ANOVA test* are not satisfied. Hence we can not conduct ANOVA test for comparing means average time across permit types.

* : You might need to login and enroll into the course to access the link

Conclusion:

In this document, one 1-sample test for population proportions and one 2-sample test for population means were described. One ANOVA test was considered but the data did not meet the conditions required.

Author	Revision History Version / Date	Comments
Aparna Shastry	0.1/03.03.2018	Initial Draft
	0.2/0.3.05.2018	Images and numbers changed because of some more data cleaning