CS 306 - Data Analysis and Visualization

Assignment 1

Kaushal Patel 201501219

1 Problem Description

Parametric testing based on t-test requires three assumptions:

1. Assumption of Normality
2. Homogeneity of Variance
3. Data Independence

These are required so that the sampling distribution of t follows the theoretical t-distribution with the corresponding degree of freedom.The goal is to verify the role of first two assumption. To that end, obtain the sampling distribution of t in four cases, and analyze the role of the said assumptions:

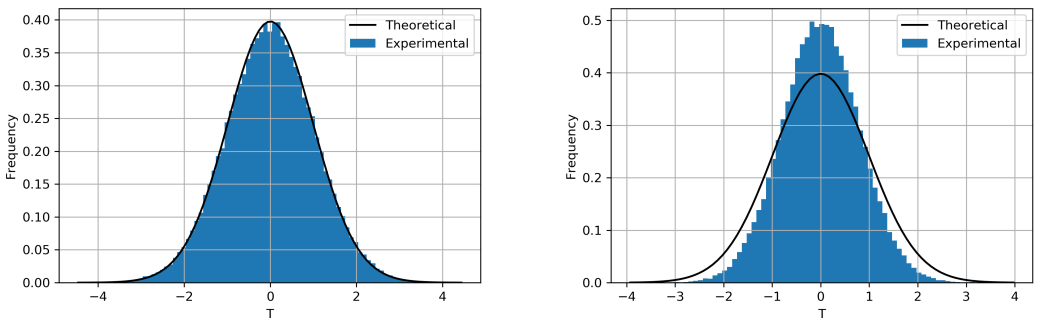
1. Normal samples with similar variance
2. Non-Normal samples with similar variance
3. Normal samples with very different variance
4. Non-Normal samples with very different variance

Your analysis should provide answers for the following:

1. is sample normality or population normality required for t-test?
2. is homogeneity of variance necessary?
3. does your answer to previous question depend on whether the sample sizes are equal or not?
4. what are the implications if one performs t-test and one or both assumptions are violated.

2 Assumptions of Normality

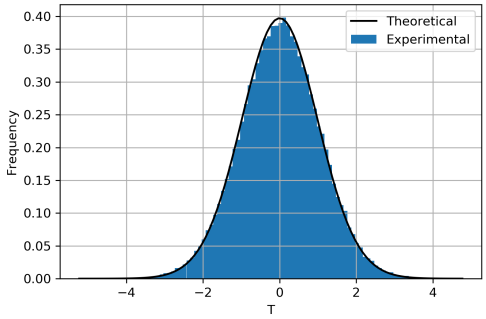
For the verification of this assumption following experimental design is implemented:
Total 10^5 times samples are taken for this experiment.
Sample size taken = 40



(a) $n_1 = 40, n_2 = 40, \sigma_1 = 1, \sigma_2 = 1$ (b) $n_1 = 40, n_2 = 90, \sigma_1 = 1, \sigma_2 = 2$

Figure 1: t Score distribution for a normally distributed population

In the above experiment we have considered that the samples drawn are from the normal population. We can observe that in figure 1 we can see that our experimental result follows the t distribution.But when we take normal population but change the experiment design such that both $var1! = var2$ and $n1! = n2$ where n1 ,n2 are sample size of the samples than we can observe that the experimental results doesnt follow the T Distribution.



(a) $n_1 = 30, n_2 = 30, \sigma_1 = 9.23, \sigma_2 = 9.23$

Figure 2: t Score distribution for a uniformly distributed population

While taking non normal population(Uniform distribution) we can observe that our experimental still follows the t distribution.Looking at the above experiment we can say that population data need not be normal in order to follow t test. Taking very skewed non normal population under the balanced experiment design(Equal sample size) still our experiment result almost follows the t distribution.

3 Homogeneity of Variance and effect of Sample size

Theoretically, t_{pooled} scores have the following distribution:

$$t_{pooled} = ct_f \tag{1}$$

where,

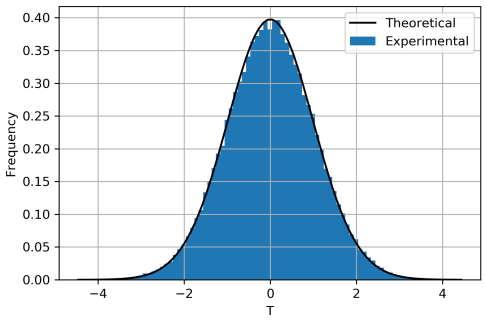
$$c = \sqrt{\frac{(n_1 + n_2 - 2)(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})}{(\frac{1}{n_1} + \frac{1}{n_2})((n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2)}} \tag{2}$$

$$\tag{3}$$

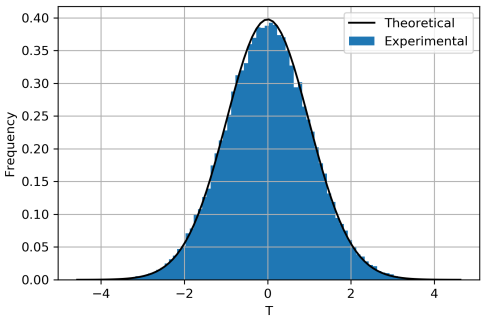
in order to follow the t-test we want c=1 which can be achieved by following two cases :

$$n_1 = n_2$$
$$\sigma_1 = \sigma_2$$

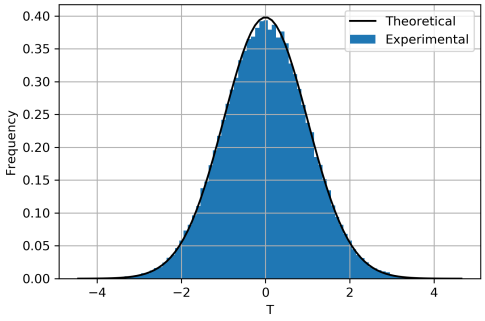
To verify above two possible cases : following are the graphs.



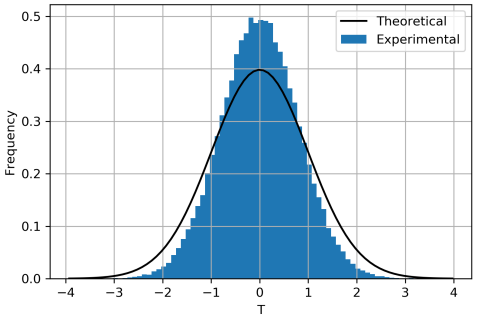
(a) $n_1 = 40, n_2 = 40, \sigma_1 = 1, \sigma_2 = 1$



(b) $n_1 = 40, n_2 = 40, \sigma_1 = 1, \sigma_2 = 2$

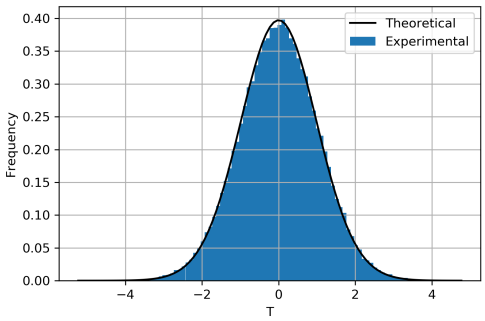


(c) $n_1 = 30, n_2 = 90, \sigma_1 = 1, \sigma_2 = 1$

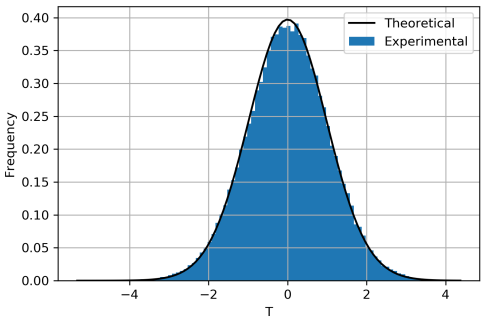


(d) $n_1 = 40, n_2 = 90, \sigma_1 = 1, \sigma_2 = 2$

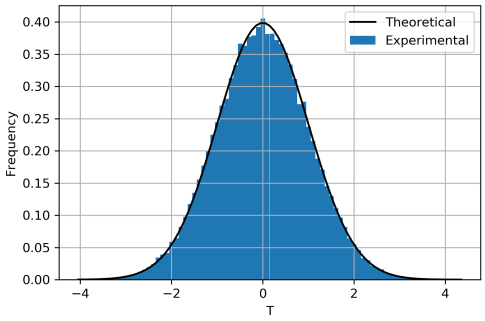
Figure 3: t Score distribution for a normally distributed population



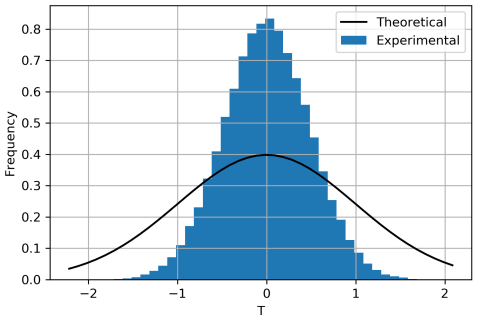
(a) $n_1 = 30, n_2 = 30, \sigma_1 = 9.23, \sigma_2 = 9.23$



(b) $n_1 = 30, n_2 = 30, \sigma_1 = 5.77, \sigma_2 = 11.54$



(c) $n_1 = 26, n_2 = 260, \sigma_1 = 2.88, \sigma_2 = 2.88$



(d) $n_1 = 26, n_2 = 210, \sigma_1 = 2.88, \sigma_2 = 8.66$

Figure 4: t Score distribution for a uniformly distributed population

From the above figures we can observe that whenever the sample size or the population variance are same than our experiment follows the t-test and in this cases the value of c=1.But when both the conditions are violated then the t distribution is not followed.
Results for the four different cases:

1. **Similar variances and similar sample sizes**: The t_{pooled} distribution closely follows the theoretical t
2. **Similar variances and different sample sizes**: The t_{pooled} distribution closely follows the theoretical t
3. **Different variances and similar sample sizes**: The t_{pooled} distribution closely follows the theoretical t
4. **Different variances and different sample sizes**: The t_{pooled} distribution does not match closely with the theoretical t

4 Conclusion

The Assumptions of Normality for t-test is incorrect. T-test or any parametric tests doesn't need to assume normality of sample nor for the underlying population from which the samples are taken. Parametric tests assume that the sampling distribution of the statistic here mean should be normally distributed. Sampling distributions approach normality as sample size increases as shown by the Central Limit Theorem.

The assumption of homogeneity of variance is an assumption of the t-test stating that all comparison groups have the same variance. The independent samples t-test utilize the t which is generally robust to violations of the assumption of Homogeneity of variance as long as sample sizes are equal.(Sample variance can be different but population variance must be same.) Thus assumption of homogeneity of variance can be discarded if we are sure that sample size of both the samples remain equal.

If sample sizes are vastly unequal and homogeneity of variance is violated, then the T statistic will be biased when large sample variances are associated with small group sizes. Here the significance level will cause the null hypothesis to be falsely rejected as the *alpha* will be underestimated. Type 1 errors will be caused as the fraction of outcomes for which the null hypothesis is actually true, but is rejected by the t test.

To be on safer side if homogeneity of variance is not satisfied then check the data , outliers in data. Also always target the balanced experimental design that is sample size of both the samples must be equal $n1 = n2$