Evaluation of Student's two-sample t-test in Qlik analytics leveraging comparison to R language

IGOR ALCANTARA* , ALANE MIGUELIS , PRISCILA RUBIM , MARK MEERSMAN

Data Science, IPC Global, United States

*Corresponding author: Igor Alcantara (igor.alcantara@ipc-global.com)

ABSTRACT

Objective

The statistical determination of a large or small difference between two groups is not based on an absolute standard, but is rather an evaluation of the probability of an event.[1,2] In the field of medical research, it is common to use statistical software for descriptive statistics as well as to perform statistical tests.[3] However, most software provides ready-to-use functions, and the researchers have almost no information as to how those statistical tests are calculated inside those functions. This article evaluates the accuracy of two-sample Student's t-test using Qlik analytics software. The gold standard used for this evaluation is the set of standard t-test functions available in R software, a widely used, robust, and reliable statistical software.[5–7]

Materials and Methods

The tests performed in this evaluation used a subset of Framingham heart study data. The dataset contains data on 4,434 anonymous participants, collected in three periods apart from each other by 6 years from 1956 to 1968. Five t-tests with 2 scenarios each were performed in Qlik analytics and in R and the results compared.

Results

In general, the results for multiple statistics obtained in Qlik analytics match the ones found in R for multiple scenarios: small and large sample sizes, small and large p-values, assuming and not assuming equal variance.

Discussion

Although Qlik analytics matches all statistics for t-test found in R, the p-value only matches up to four decimal points, which is concluded to be enough for testing hypothesis since the conventional levels of significance do not go lower than 0.1.

Conclusion

This research concluded that Qlik analytics can be used for two-sample t-tests in multiple scenarios.

Keywords: Qlik, t-test, r language, Framingham.

INTRODUCTION

Background and Significance

In 1908, the English statistician, chemist and brewer William Sealy Gosset, better known as Student, published his famous works entitled "The Probable Error of a Mean" and "Probable Error of a Correlation Coefficient".[8,9] His articles inferred a distribution of a statistic now known as Student's t, and introduced small sample estimation by a family of t-distributions.[10,11] From this work by Gosset, Ronald Fisher derived the exact distribution of the correlation coefficient[12].

In 1912, Ronald Fisher wrote to Gosset about an error in Gosset's formula for standard deviation, which is used in z-test calculation, he has found in Gosset's paper. Fisher said that the total sample size should be reduced by one and sent the proof of this statement. With this, Fisher also changed the nomenclature from z-test to t-test.

Statistical tests, including the t-test, have been applied in a variety of settings where sampling is performed and experimenters are interested in testing population parameters based on sampling information.[13] The most commonly used t-tests are: Comparison of the mean of a single sample (one-sample t-test); tests for two related samples (two-sample t-test); tests for correlation and regression coefficients against a hypothetical value, which is usually zero (paired t-test).[14,15]. In the case of the present work, we used two-sample t-test, which is available in both R programming language and Qlik analytics software.

This study uses a subset of the Framingham Heart study data. This data is familiar to many medical researchers, as it is described in the Data Records section. Although we are using medical data, this is not a medical study. The objective of this article is to evaluate how well Qlik software calculates the various statistics related to a two-sample t-test.

Qlik is a data analytics software company founded in 1993 in Lund, Sweeden.[16] It is widely used as a Business Intelligence[17] analytics platform.[18] According to the Gartner Magic Quadrant for Business Intelligence Platforms,[19] Qlik is a leader in the market, being user-friendly and requiring almost no software programming skills.

Objective

Researchers often need to use statistical software to run analyses, but this task requires programming skills that take time to be mastered.[20] Qlik is a platform that is already in use by researchers for data visualization[21,22] and preparation.[23,24] This fact motivated this study with the objective of evaluating the accuracy of Qlik's functions for the t-test. This study assesses the reliability of Qlik as s tool for hypothesis testing.

MATERIALS AND METHODS

Methods

The accuracy of the Qlik platform when performing the two-sample t-test was measured by comparing the test results performed in Qlik with the same tests performed in R. The tasks were split among the different researchers where the group performing the tests in R did not have contact with the group performing the tests in Qlik until the final results were compared.

Five tests were performed. In the first test, the total cholesterol in males was compared to the total cholesterol in females to measure if the difference is statistically significant. In the second test, the systolic blood pressure of current smokers was compared to the systolic blood pressure of non-current smokers. In the third test, the total cholesterol was measured against two groups: diabetic and non-diabetic participants. In the fourth test an inclusion criteria was added: diabetic patients 50 years old or younger who at some point died of any cause. The mean systolic blood pressure was measured in two groups: men and women. Finally, in the last test, the systolic blood pressure of diabetic current smokers was compared to the systolic blood pressure of diabetic non or former smokers. For all tests, two scenarios were considered: equal variance assumed, and equal variance not assumed. Therefore, in total, 10 evaluations were performed.

The main objective of this article was to assess if Qlik is a reliable software to perform two-sample t-tests, which is why multiple tests were performed.

Two quantitative variables were used for this study, as described in the Data Records section of this article: Total Cholesterol (mg/dL) and Systolic Blood Pressure (mmHg). The Figure 1 shows the distribution plot for Total Cholesterol for all observations. Although the shape of the distribution seems to be normal,[25] the measure of Skewness demonstrates otherwise.[26] The measure of Kurtosis[27] shows the presence of some outliers in the data. These numbers,[28] as measured by Qlik analytics software, can be seen in table 1. The same table also shows other statistics for Total Cholesterol, including the Total Change Mean Square Error (MSE) for this metric among all those patients. The number found for MSE is small, indicating overall very little change in this metric over time.

Statistic	R Value	Qlik Value
n	11,627	11,627
mean	241.2	241.2
median	238	238
standard deviation	45.37	45.37
Kurtosis	3.38	3.38
Skewness	0.82	0.82
Total Change MSE	0.54	0.54

Table 1. Summary Statistics for Total Cholesterol.

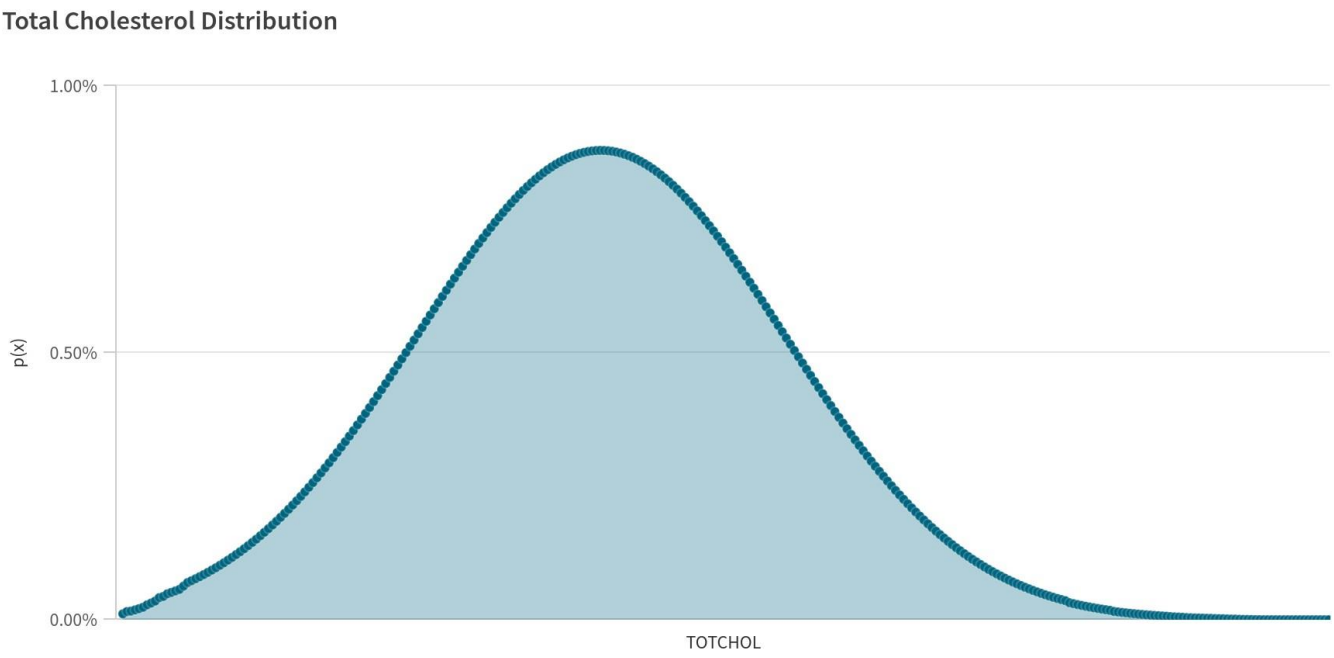

Figure 1. Distribution plot for Total Cholesterol. Graph was generated by Qlik software

Figure 2 shows the distribution plot[15] for Systolic Blood Pressure for all observations. Similar to what is observed for Total Cholesterol, the shape of the distribution seems to befg6 normal[29], but the measure of Skewness[30] demonstrates it not to be gaussian.[31] The measure of Kurtosis[32] shows the presence of very few outliers in the data. These numbers as measured by Qlik analytics software can be seen in Table 2. The same table also shows other statistics for Systolic Blood Pressure,

including the Total Change Mean Square Error (MSE) for this metric among all those patients. The number found for MSE is small, indicating overall very little change in this metric over time.

Statistic	R Value	Qlik Value
n	11,627	11,627
mean	136.3	136.3
median	132	132
standard deviation	22.8	22.8
Kurtosis	1.37	1.37
Skewness	0.94	0.94
Total Change MSE	0.3	0.3

Table 2. Summary Statistics for Systolic Blood Pressure.

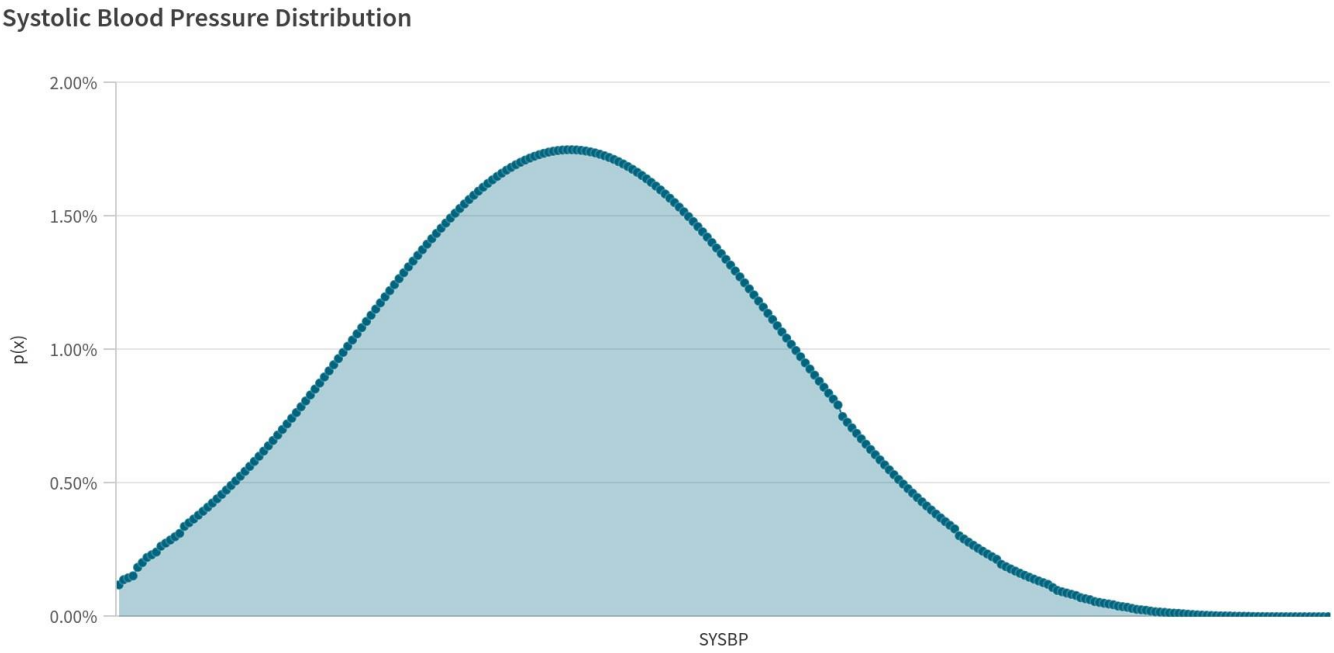

Figure 2. Distribution plot for Systolic Blood Pressure. Graph was generated by Qlik software

The shape and skewness of the data validates the use of a two-sample t-test.[33] In each one of the tests performed, the data was split into two independent groups: Male/Female, Smokers/Non-Smokers, and Diabetics/Non-Diabetics. For each group, one participant might have its measures collected in different periods but the individual change of the individual metrics is minimum. As described in Table 1 and in Table 2, the Mean Square Error change in Total Cholesterol and Systolic Blood Presssure are 0.54 mg/dL and 0.3 mmHg respectfully. We also considered the fact that this paper did not intend to make any medical claims about correlation between participants' conditions but to test whether the Qlik calculation of the selected types of t-test is accurate in comparison to R. For those reasons, a two-sample t-test[1] was selected for this study.

In order to demonstrate Qlik's t-test function accuracy in different scenarios, multiple tests were performed. Tests 1 and 2 were selected to evaluate large sample sizes where the mean of one group is significantly different than the other. Test 3 evaluated Qlik's accuracy for non-significant mean difference between groups and large p-value. Test 4 analyzes t-test

performance against small sample sizes. In addition, Test 5 was performed to test the precision of p-value between the two technologies: Qlik and R.

Data Records

This article used a well-known dataset in the medical research field.[34] The Framingham Heart Study is a longitudinal long term study study of the etiology of cardiovascular disease among the population of Framingham, Massachusetts.[35] The data produced by the Framingham Heart Study is largely used by the medical research community,[36] for analysis of several disease risk factors[37] over multiple generations.[38] It is a good use case to test accuracy of statistical software.[39]

Framingham Study data

The dataset used for this article was provided by the U.S. National Institutes of Health though the National Heart, Lung and Blood Institute Biologic Specimen and Data Repository (BioLINCC).[40] The dataset provided by BioLINCC contains laboratory, clinical and demographics data on 4,434 participants. The data is a subset of the Framingham Heart Study and it was collected in three periods apart from each other by 6 years from 1956 to 1968.

Each participant was followed for a total of 24 years after the final period for the outcome of some health-related events such as Angina Pectoris, Myocardial Infarction, Atherothrombotic Infarction or Cerebral Hemorrhage (Stroke) or death.

The dataset provided was anonymized prior to delivery to guarantee the participants' privacy. The study is a longitudinal study. Each participant has 1, 2 or 3 observations depending on the number of medical appointments the subject attended. In the data provided, each period is identified as either 1, 2 or 3. As a total, there are 11,627 observations for the 4,434 participants.

Data used in the study

The variables used in this study are described in Table 3. For categorical variables (Sex, CurrSmoke, Diabetes, Death) the Range/Count column shows the sample size for each category (e.g. 5022 men and 6605 women), and for numeric values the Range/Count column shows the minimum-maximum range encounter for that variable.

Variable	Description	Units	Range or Count
Sex	Participant Sex	1=Man | 2=Woman	n=5022 | n=6605
CurrSmoke	Current cigarette smoking at exam	0=Not current smoker | 1=Current smoker	n=6598 | n=5029
Diabetes	Diabetic according to glucode of 200mg/dL or more	0=Not a diabetic | 1=Diabetic	n=11097 | n=530
Age	Age at exam (years)	n/a	32-81
Death	Death of any cause	0=No | 1=Yes	n=2884 | n=1550
SysBP	Systolic Blood Pressure (mmHg)	n/a	83.5-295
TotChol	Serum Total Cholesterol (mg/dL)	n/a	107-696

Table 3. List of variables used in the study.

RESULTS

Test 1: Total Cholesterol by Sex

The first test split the data into two groups: male, and female participants and it tested if the difference in Total Cholesterol is significantly different between genders. According to what was calculated by native R functions, the mean of Total Cholesterol for all participants, excluding nulls, is 241.1624 with a standard deviation of 45.36803. The mean Total Cholesterol for males is 234.2818 with a standard deviation of 42.41163. The mean Total Cholesterol for females is 246.5278 with a standard deviation of 46.85322.

The calculations of Mean and Standard Deviation in Qlik found the same results as the R code. The observed mean of Total Cholesterol is larger in women than in men. A two-tailed two-sample t-test was conducted to evaluate if this difference is statistically significant. The null hypothesis is that there is no significant difference, and the alternative hypothesis is that there is significant difference.

The first scenario assumed equal variance and a t-test was performed for a 95% confidence level and alpha equal to 5%. With 11,216 degrees of freedom, the t-value found was 14.313. The confidence interval (CI) is 10.56896 to 13.92314. The p-value is < 2.2e-16. Therefore, the null hypothesis is rejected.

A second scenario not assuming equal variance was performed for a 95% confidence level and alpha equal to 5%. With 10,970 degrees of freedom, the t-value found was 14.49. The confidence interval is 10.58944 to 13.90267. The p-value is < 2.2e-16. Therefore, the null hypothesis is rejected.

The statistics found in Qlik are almost the same found in R as described in Table 4. The p-value calculated in R is presented in a different notation than the one in Qlik, which will be discussed further.

Equal Variance	Statistic	R result	Qlik result
	mean Total Chol	241.624	241.162
	standard deviation Total Chol	45.36803	45.36803
	mean Total Chol for males	234.2818	234.2818
	standard deviation Total Chol for males	42.41163	42.41163
	mean Total Chol for females	246.5278	246.5278
	standard deviation Total Chol for females	46.85322	46.85322
Yes	degrees of freedom	11216	11216
Yes	t-value	14.313	14.1313
Yes	Lower CI	10.56896	10.56896
Yes	Upper CI	13.92314	13.92314
Yes	p-value	< 2.2e-16	8.2e-15
No	degrees of freedom	10970	10970
No	t-value	14.49	14.49
No	Lower CI	10.58944	10.58944
No	Upper CI	13.90267	13.90267
No	p-value	< 2.2e-16	1.9e-15

Table 4. Comparison of statistics found in R and in Qlik for the 1st Test.

Test 2: Systolic Blood Pressure by Smoking Status

The second test splits the data into two groups: current smokers and non-smokers, and it tested if the difference in Systolic Blood Pressure is significantly different between smoking status groups. According to what was calculated by native R functions, the mean of Systolic Blood Pressure for all participants, excluding nulls, is 136.3241 with a standard deviation of 22.79862. The mean Systolic blood pressure for current smokers is 132.6484 with a standard deviation of 21.74775. The mean Systolic blood pressure for non-smokers is 139.1257 with a standard deviation of 23.18158.

The calculations of Mean and Standard Deviation in Qlik found the same results as the R code. The observed mean Systolic Blood Pressure for non-smokers is larger than the one for current smokers. A two-tailed two-sample t-test was conducted to evaluate if this difference is statistically significant. The null hypothesis is that there is no significant difference, and the alternative hypothesis is that there is significant difference.

The first scenario assumed equal variance and a t-test was performed for a 95% confidence level and alpha equal to 5%. With 11,625 degrees of freedom, the t-test found was 15.329. The confidence interval is 5.649030 to 7.305532. The p-value is <2.2e-16. Therefore, the null hypothesis is rejected.

A second scenario not assuming equal variance was performed for a 95% confidence level and alpha equal to 5%. With 11,140 degrees of freedom, the t-value found was 15.462. The confidence interval is 5.656123 to 7.298439. The p-value is < 2.2e-16. Therefore, the null hypothesis is rejected.

The statistics found in Qlik are the same found in R as described in Table 5. The p-value calculated in R is slightly different than the one found in Qlik, which will be discussed further.

Equal Variance	Statistic	R result	Qlik result
	mean SysBP	136.3241	136.3241
	standard deviation SysBP	22.7962	22.7962
	mean SysBP for smokers	132.6484	132.6484
	standard deviation SysBP for smokers	21.74775	21.74775
	mean SysBP for non-smokers	139.1257	139.1257
	standard deviation SysBP for non-smokers	23.18158	23.18158
Yes	degrees of freedom	11625	11625
Yes	t-value	15.329	15.329
Yes	Lower CI	5.649030	5.649030
Yes	Upper CI	7.303109	7.305532
Yes	p-value	< 2.2e-16	5.3e-15
No	degrees of freedom	11140	11140
No	t-value	15.462	15.462
No	Lower CI	5.656123	5.626123
No	Upper CI	7.298439	7.298439
No	p-value	< 2.2e-16	3.7e-15

Table 5. Comparison of statistics found in R and in Qlik for the 2nd Test.

Test 3: Total Cholesterol by presence of Diabetes

The third test split the data into two groups: diabetic, and non-diabetic participants and it tested if the difference in Total Cholesterol is significantly different between diabetic diagnoses.According to what was calculated by native R functions, the mean of Total Cholesterol for all participants, excluding nulls, is 241.1624 with a standard deviation of 45.36803. The mean Total Cholesterol for diabetic participants is is 241.8526 with a standard deviation of 53.56868. The mean Total Cholesterol for non-diabetic participants is 241.1301 with a standard deviation of 44.94987.

The calculations of Mean and Standard Deviation in Qlik found the same results as the R code. The observed mean of Total Cholesterol in diabetic participants is slightly larger than the value found for non-diabetic participants. A Two-Tailed Sample t-test was performed to evaluate if this difference is statistically significant. The null hypothesis is that there is no significant difference, and the alternative hypothesis is that there is significant difference.

The first scenario assumed equal variance and a t-test was performed for a 95% confidence level and alpha equal to 5%. With 11,216 degrees of freedom, the t-value found was 0.34873. The confidence interval is -3.338669 to 4.783676. The p-value is 0.7273. Therefore, we fail to reject the null hypothesis.

A second scenario not assuming equal variance was performed for a 95% confidence level and alpha equal to 5%. With 534.57 degrees of freedom, the t-value found was 0.29733. The confidence interval is -4.051012 to 5.496020. The p-value is 0.7663. Therefore, we fail to reject the null hypothesis.

The statistics found in Qlik are the same found in R as described in Table 6.

Equal Variance	Statistic	R result	Qlik Result
	mean Total Chol	241.624	241.162
	standard deviation Total Chol	45.36803	45.36803
	mean Total Chol for diabetic	241.8526	241.8526
	standard deviation Total Chol for diabetic	53.56868	53.56868
	mean Total Chol for non-diabetic	241.1301	241.1301
	standard deviation Total Chol for non-diabetic	44.94987	44.94987
Yes	degrees of freedom	11216	11216
Yes	t-value	0.34873	0.34873
Yes	Lower CI	-3.338669	-3.338668
Yes	Upper CI	4.783676	4.783676
Yes	p-value	0.7273	0.7273
No	degrees of freedom	534.57	534.57
No	t-value	0.29733	0.29733
No	Lower CI	-4.051012	-4.051012
No	Upper CI	5.496020	5.496020
No	p-value	0.7663	0.7663

Table 6. Comparison of statistics found in R and in Qlik for the 3rd Test.

Test 4: Systolic Blood Pressure by Sex for younger diabetic participants with recorded death

The fourth test filtered the data with specific inclusion criteria (described below) and it splits the subset of data into two groups: males and females. It also tested if the difference in Systolic Blood Pressure is significantly different between genders.

Inclusion criteria:

- Age <= 50 years old
- Diabetic
- Death recorded

According to what was calculated by native R functions, the mean Systolic blood pressure for the included participants, excluding nulls, is 138.0333, with a standard deviation of 27.0905. The mean Systolic blood pressure for the male included participants is 134.9231, with a standard deviation of 23.11822.

The calculations of Mean and Standard Deviation in Qlik found the same result as the R code. The observed mean systolic blood pressure for the male included participants is higher than the measure found for the female included participants. A two-tailed two-sample t-test was conducted to evaluate if this difference is significant. The null hypothesis is that there is no significant difference, and the alternative hypothesis is that there is significant difference.

The first scenario assumed equal variance and a t-test was performed for a 95% confidence level and alpha equal to 5%. With 28 degrees of freedom, the t-test found was 0.54318. The confidence interval is -15.20995 to 26.18732. The p-value is 0.5913. Therefore, we fail to reject the null hypothesis.

A second scenario not assuming equal variance was performed for a 95% confidence level and alpha equal to 5%. With 27.998 degrees of freedom, the t-value found was 0.56324. The confidence interval is -14.47292 to 25.45030. The p-value is 0.5778. Therefore, we fail to reject the null hypothesis.

The statistics found in Qlik are the same found in R as described in Table 8.

Equal Variance	Statistic	R result	Qlik result
	mean SysBP	138.0333	138.0333
	standard deviation SysBP	27.0905	27.0905
	mean SysBP for males	134.9231	134.9231
	standard deviation SysBP for males	23.11822	23.11822
	mean SysBP for females	140.4118	140.4118
	standard deviation SysBP for females	30.25684	30.25684
Yes	degrees of freedom	28	28
Yes	t-value	0.54318	0.54318
Yes	Lower CI	-15.20995	-15.20995
Yes	Upper CI	26.18732	26.18732
Yes	p-value	0.5913	0.5913
No	degrees of freedom	27.998	27.998
No	t-value	0.56324	0.56324
No	Lower CI	-14.47292	-14.47292
No	Upper CI	25.45030	25.45030
No	p-value	0.5778	0.5778

Table 7. Comparison of statistics found in R and in Qlik for the 4th Test.

Test 5: Systolic Blood Pressure for Diabetic participants by Smoking Status

The fifth test filtered the data for diabetic participants only, and it split the subset of data into two groups: Current Smokers and Non-Smokers, and it tested if the difference in Systolic Blood Pressure is significantly different between those groups. According to what was calculated by native R functions, the mean Systolic Blood Pressure for diabetic participants, excluding nulls, is 151.417, with a standard deviation of 27.59397. The mean Systolic blood pressure for the current smokers diabetic participants is 145.8842, with a standard deviation of 28.00166. The mean Systolic blood pressure for the non-smokers diabetic participants is 154.1912, with a standard deviation of 27.0027.

The calculations of Mean and Standard Deviation in Qlik found the same result as the R code. The observed mean systolic blood pressure for the non-current smokers diabetic participants is higher than the measure found for the current smokers diabetic participants. A two-tailed two-sample t-test was conducted to evaluate if this difference is significant. The null hypothesis is that there is no significant difference, and the alternative hypothesis is that there is significant difference. The first scenario assumed equal variance and a t-test was performed for a 95% confidence level and alpha equal to 5%. With 528 degrees of freedom, the t-test found was 3.299. The confidence interval is 3.360479 to 13.253596. The p-value is 0.001036 (0.0010355833153088). Therefore, we reject the null hypothesis.

A second scenario not assuming equal variance was performed for a 95% confidence level and alpha equal to 5%. With 341.29 degrees of freedom, the t-value found was 3.2594. The confidence interval is 3.294053 to 13.320021. The p-value is 0.001229 (0.0012287762526965). Therefore, we reject the null hypothesis.

The statistics found in Qlik are the same found in R as described in table 8.

Equal Variance	Statistic	R result	Qlik result
	mean diabetic SysBP	151.417	151.417
	standard deviation diabetic SysBP	27.59391	27.59391
	mean SysBP for diabetic current smokers	145.8842	145.8842
	standard deviation SysBP for diabetic current smokers	28.00166	28.00166
	mean SysBP for diabetic non-smokers	154.1912	154.1912
	standard deviation SysBP for non-smokers	27.0027	27.0027
Yes	degrees of freedom	528	528
Yes	t-value	3.299	3.299
Yes	Lower CI	3.360479	3.360479
Yes	Upper CI	13.253596	13.253596
Yes	p-value	0.001036	0.001036
Yes	p-value (16f)	0.0010355833153088	0.0010355833153106
No	degrees of freedom	341.29	341.29
No	t-value	3.2594	3.2594
No	Lower CI	3.294053	3.294053
No	Upper CI	13.320021	13.320021
No	p-value	0.001229	0.001229
Yes	p-value (16f)	0.0012287762526965	0.0012288758442427

Table 8. Comparison of statistics found in R and in Qlik for the 5th Test.

DISCUSSION

This study performed five two-sample t-tests where each test was performed in two scenarios: considering equal variance and not assuming equal variance, giving a total of 10 t-tests. In 6 of the 10 tests, the p-value was smaller than the significance level and the null hypothesis was rejected. In 4 of 10 tests, the p-value was above the significance level and the null hypothesis was not rejected.

Different variables were used for the tests, as described in Table 3. An inclusion criteria was introduced for the fourth group of tests to verify how well Qlik performs t-tests for small samples. A subset only with diabetes patients was selected for a fifth group of tests to measure the precision of p-value found in both technologies. Qlik and R were compared against 7 different statistics:

- Mean
- Standard deviation
- Degrees of freedom
- t-value
- Lower value for Confidence Interval
- Upper value for Confidence Interval
- p-value

For all of the scenarios presented, Qlik standard statistical functions found very similar results to R native statistic functions. In cases where the p-value was too small, the scientific notation used by R and Qlik are different, making it more difficult to compare exact results.

According to the R documentation,[41] if the p-value calculated is less than 2.2e-15, R interface simply shows that the p-value is smaller than that threshold. Qlik is more precise when showing p-values, allowing researchers to specify how many decimal places are displayed. When comparing the p-value found in R with the p-value found in Qlik there is a small difference. The values match for up to some decimal places but as we add more precision some differences are seen. For example, for the second t-test, when assuming equal variance, the p-value is flagged as being less than 2.2e-16. The p-value found in Qlik is 5.3e-15, which is not smaller than 2.2e-16.
There are many reasons that could explain these differences for highly precise numbers, for example, how each software rounds the parameters used to calculate the p-value.[42] To understand this difference, we included a fifth test, where the p-value was not smaller than R lower limit of 2.2e-15. Analyzing Test 5, when assuming equal variance, both R and Qlik find a p-value of 3.299. When expanded to more precision, Qlik matches R up to 13 decimal places (R shows 0.0010355833153088 and Qlik shows 0.0010355833153106). Test 5 not assuming equal variance, both R and Qlik find a p-value of 0.001229. When adding more precision, Qlik matches R up to 6 decimal places (R shows 0.0012287762526965 and Qlik shows 0.0012288758442427).

For p-values with up to 6 decimal places, Qlik found the exact same results as R. In case the p-value is too small, a more accurate alternative is to use the t-value and compare it against the value found in the t reference table.[43] T-value was consistent between R and Qlik for all of the tests performed.

CONCLUSION

We conclude that Qlik is a reliable software to perform two-sample t-tests, both for large and for small sample sizes. The difference found in the p-value was only visible for very small p-values which do not change the conclusion of the test in whether to reject or fail to reject the null hypothesis.[44] In addition, the t statistic found was always consistent, giving the researcher using Qlik the ability to compare the findings to the reference value found in the reference table.[45]

Future work

In future work, we intend to expand the evaluation of Qlik's statistical functions to other hypothesis tests: one-sample t-test, and chi-2. We also intend to expand the current work evaluating two-sample t-test using different datasets.

Code Availability

This study uses two technologies: R and Qlik. As an Integrated Development Environment (IDE) for R, we use R Studio. R can be downloaded at https://cran.r-project.org/bin/windows/base/, and R Studio at https://www.rstudio.com/products/rstudio/download/. Both R and R studio are free. Qlik software does not require installation if used the SaaS version. Qlik is not free but there is a trial version it can be accessed at https://www.qlik. com/. In order to run the Qlik application, one can simply download the qvf file provided at the GitHub repository and upload to Qlik platform.

The data and source code used in this study is available in a GitHub repository. The R Code is fully available. Only one package is needed to execute the code: Dplyr, which can be easily installed in R Studio via the Tools menu.

The Qlik application file is provided as a ".qvf" file, which can be imported to the same or ealier version of Qlik Sense and Qlik SaaS software. The qvf file contains all source code and user interface. In addition, a file with the load script is available. For someone with minimal Qlik knowledge it should be simple to reproduce the code.

The version of R language used is 4.1.2. The code was written and executed in R Studio version 2022.02.03. The Qlik product used for this article is Qlik SaaS version 12.1438.0.

The code and results can be also reproduced in other products of the Qlik family, as long as they are either Qlik Sense or Qlik SaaS. These codes were not tested in a QlikView environment. The data is available for download in the GitHub repository. This file is named as "Framingham_Data". The data file is the same provided by BioLINCC. The data is anonymous and contains no identifiable participant information. The GitHub repository can be accessed by the following link: https://github.com/ipc-data-science/qlik-r-t-test-2-sample.

Acknowledgments

This study was performed using data provided by the Biologic Specimen and Data Repository (BioLINCC) from the National Heart, Lung, and Blood Institute. We would also like to thank Dr. Louise Jones. PhD., director of GME Research & Scholarly Activity at Northeast Georgia Health System for her support and to engage our group in multiple important medical research.

Author information

These authors contributed equally: Igor Alcantara, Alane Miguelis, and Priscila Rubim.

Authors and Affiliations

IPC Global, Data Science

Igor Alcantara, Alane Miguelis, Dr. Priscila Rubim, PhD, and Mark Meersman.

Contributions

Igor Alcantara, Alane Miguelis, and Priscila Rubim designed the study, acquired the data, defined the methodology, review of the literature, performed the statistical tests in R and in Qlik, created the project code repository, and drafted the manuscript. Mark Meersman secured resources and funding, and performed the final review of the manuscript.

Corresponding author

Igor Alcantara (igor.alcantara@ipc-global.com)

Competing interests

IPC Global is a partner of Qlik. This could be identified as an employment competing interest. The authors declare all the necessary effort was done to reduce bias in the analysis performed.

Funding information

This research is self-funded by IPC Global. No external funding was used.

References

1. Kim, T. K. T test as a parametric statistic, 10.4097/kjae.2015.68.6.540 (2015).
2. de Verdier, C. H. What is the quality of quality control procedures?? *Scand. J. Clin. Lab. Investig.* 41, 1–14, 10.3109/00365518109092008 (1981).
3. Jiaxi, L. The application and research of t-test in medicine. 321–323, 10.1109/ICNDC.2010.70 (2010).
4. Zou, H. Study on t-test method to find linear systematic error. 10.1109/MEC.2011.6026003 (2011).
5. Shedlock, C. J. & Stumpo, K. A. Data parsing in mass spectrometry imaging using r studio and cardinal: A tutorial. *J. Mass Spectrom. Adv. Clin. Lab* 23, 58–70, 10.1016/j.jmsacl.2021.12.007 (2022).
6. Davies, B. M. *et al.* Current surgical practice for multi-level degenerative cervical myelopathy: Findings from an international survey of spinal surgeons. *J. Clin. Neurosci.* 87, 84–88, 10.1016/j.jocn.2021.01.049 (2021).
7. Feng, H. *et al.* Genomic features and clinical characteristics of adolescents and young adults with cholangiocarcinoma. *Front. Oncol.* 9, 10.3389/fonc.2019.01439 (2020).
8. Student. The probable error of a mean. *Biometrila* 6, 1–25, https://doi.org/10.2307/2331554 (1908).
9. Student. Probable error of a correlation coefficien. *Biom. Trust.* 6, 302–310, https://doi.org/10.2307/2331474 (1908).
10. Box, J. F. Guinness, gosset, fisher and small samples. *Stat. Sci.* 1, 45–52, 10.1214/ss/1177013437 (1987).
11. Zabell, S. L. On student's 1908 article "the probable error of a mean". *J. Am. Stat. Assoc.* 103, 1–7, 10.1198/016214508000000030 (2008).
12. Fisher, R. A. Applications of "student's, distribution. (1925).
13. Albassam, M. & Aslam, M. Testing internal quality control of clinical laboratory data using paired t-test under uncertainty. *BioMed Res. Int.* 2021, 10.1155/2021/5527845 (2021).
14. Pandey, R. Commonly used t-tests in medical research. *J. Pract. Cardiovasc. Sci.* 1, 185, 10.4103/2395-5414.166321 (2015).
15. Skaik, Y. The bread and butter of statistical analysis "t-test": Uses and misuses. *Pak. J. Med. Sci.* 31, 1558–1559, 10.12669/pjms.316.8984 (2015).
16. Christiansen, E. The success stories you probably haven't heard of: Qlik. Available at https://oresundstartups.com/success-stories-might-missed-qlik/ (2019).
17. Luhn, H. P. A business intelligence system. *IBM J. Res. Dev.* 2, 314–319, 10.1147/rd.24.0314 (1958).
18. Yordanova, Z. Innovation development and ramp;d project management in science organizations and universities data-driven model and analysis, 10.1007/978-3-030-68133-3_1 (2021).
19. Austin Kronz, J. S. D. P. A. G., Kurt Schlegel. Gartner magic quadrant for analytics and business intelligence platforms. Available at https://www.gartner.com/doc/reprints?id=1-292LEME3&ct=220209&st=sb (2022).
20. Baker, M. Scientific computing: Code alert. *Nature* 541, 563–565, 10.1038/nj7638-563a (2017).
21. Tiozzo, B. *et al.* Biological, chemical, and nutritional food risks and food safety issues from italian online information sources: Web monitoring, content analysis, and data visualization. *J. Med. Internet Res.* 22, e23438, 10.2196/23438 (2020).
22. Reid, N. E., Johnson-Arbor, K., Smolinske, S. & Litovitz, T. 2020 webpoisoncontrol data summary. *The Am. J. Emerg. Medicine* 54, 184–195, 10.1016/j.ajem.2022.02.014 (2022).
23. Barker, R., Lombardino, J., Rasmussen, K. & Gilroy, S. Test of arabidopsis space transcriptome: A discovery environment to explore multiple plant biology spaceflight experiments. *Front. Plant Sci.* 11, 10.3389/fpls.2020.00147 (2020).
24. Bodí, M. *et al.* Automatic generation of minimum dataset and quality indicators from data collected routinely by the clinical information system in an intensive care unit. *Int. J. Med. Informatics* 145, 104327, 10.1016/j.ijmedinf.2020.104327 (2021).

25. Maltenfort, M. G. Understanding a normal distribution of data. *J. Spinal Disord. Tech.* 28, 377–378, 10.1097/BSD.0000000000000337 (2015).

26. Doane, D. P. & Seward, L. E. Measuring skewness: A forgotten statistic? *J. Stat. Educ.* 19, 10.1080/10691898.2011.11889611 (2011).

27. Decarlo, L. T. On the meaning and use of kurtosis = Available at http://dx.doi.org/10.1037/1082-989X.2.3.292/. *Psychol. Methods* 292–307 (1997).

28. Mishra, P. *et al.* Descriptive statistics and normality tests for statistical data. *Annals Cardiac Anaesth.* 22, 67, 10.4103/aca.ACA_157_18 (2019).

29. Reid, S. What is so normal about the normal distribution? *Evidence-Based Mental Heal.* 13, 100–100, 10.1136/ebmh.13.4. 100 (2010).

30. Royston, P. Which measures of skewness and kurtosis are best? *Stat. Medicine* 11, 333–343, 10.1002/sim.4780110306 (1992).

31. Hopkins, K. D. & Weeks, D. L. Tests for normality and measures of skewness and kurtosis: Their place in research reporting. *Educ. Psychol. Meas.* 50, 717–729, 10.1177/0013164490504001 (1990).

32. Fiori, A. M. & Zenga, M. Karl pearson and the origin of kurtosis. *Int. Stat. Rev.* 77, 40–50, 10.1111/j.1751-5823.2009.00076.x (2009).

33. Garren, S. & Osborne, K. Robustness of t-test based on skewness and kurtosis. *J. Adv. Math. Comput. Sci.* 102–110, 10.9734/jamcs/2021/v36i230342 (2021).

34. Andersson, C., Nayor, M., Tsao, C. W., Levy, D. & Vasan, R. S. Framingham heart study: Jacc focus seminar. *J. Am. Coll. Cardiol.* 77, 2680–2692, 10.1016/j.jacc.2021.01.059 (2021).

35. Mahmood, S. S., Levy, D., Vasan, R. S. & Wang, T. J. The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet* 383, 999–1008, 10.1016/S0140-6736(13)61752-3 (2014).

36. Peeters, A. A cardiovascular life history. a life course analysis of the original framingham heart study cohort. *Eur. Hear. J.* 23, 458–466, 10.1053/euhj.2001.2838 (2002).

37. Bhargava, A. A longitudinal analysis of the risk factors for diabetes and coronary heart disease in the framingham offspring study. *Popul. Heal. Metrics* 1, 3, 10.1186/1478-7954-1-3 (2003).

38. Yashin, A. I. *et al.* "predicting" parental longevity from offspring endophenotypes: Data from the long life family study (llfs). *Mech. Ageing Dev.* 131, 215–222, 10.1016/j.mad.2010.02.001 (2010).

39. Ignjatovic, A., Stojanovic, M., Milosevic, Z. & Apostolovic, M. A. Progress of statistical analysis in biomedical research through the historical review of the development of the framingham score. *Ir. J. Med. Sci. (1971 -)* 187, 639–645, 10.1007/s11845-017-1718-5 (2018).

40. Nhlbi biologic specimen and data repositories. Available at https://biolincc.nhlbi.nih.gov (2022).

41. R documentation for t-test function. Available at https://search.r-project.org/R/refmans/stats/html/prop.test.html (2022).

42. Blackstone, E. H. Rounding numbers. *The J. Thorac. Cardiovasc. Surg.* 152, 1481–1483, 10.1016/j.jtcvs.2016.09.003 (2016).

43. Posten, H. O. *Robustness of the Two-Sample T-Test* (Springer Netherlands, 1984).

44. Krzywinski, M. & Altman, N. Significance, p values and t-tests. *Nat. Methods* 10, 1041–1042, 10.1038/nmeth.2698 (2013).

45. Dawson, R. J. M. Turning the tables: A t-table for today. *J. Stat. Educ.* 5, 10.1080/10691898.1997.11910530 (1997).