**CURE in Traits and Species Distributions**

**Statistical Analysis Module**

**Introduction**

Welcome to the statistical analysis module. Congratulations! You have done a lot of work to get to this point. Statistical analysis provides us with opportunity to test hypotheses and communicate findings. To conduct statistical analyses for this project, we will be using Jeffery’s Amazing Statistics Program (JASP). JASP is an open-source software for statistical analysis and data visualization that was developed by Love *et al.* in 2019. JASP employes a graphical user interface (GUI) to execute statistical analyses using the R programming language. Before starting this module, please go to <https://jasp-stats.org/download/> and download the appropriate JASP installer for your operating system.

**Part 1. Brief Introduction to Biological Statistics**

1. Types of Variables

* **Independent Variables (X)** - variables we assume will cause a change in a dependent variable. Also referred to as **predictor** or **explanatory** variables.
* **Dependent Variables (Y) –** variables we assume to be affected by a dependent variable. Also referred to as **response** variables**.**

1. Assumptions
   * Tests for **normality** use the mean of your data set to determine the *distribution* of your data in relation to the mean. A dataset with a normal or Gaussian distribution will be represented with a bell-shaped curve when plot using a histogram suggesting a majority of the data is close to the mean. A dataset lacking normality will be skewed with a majority of data points away from the mean.
   * **Homoscedasticity** describesdata with the same standard deviation across all groups. Standard deviation is a measure of variance in square units. Data that have an unequal standard deviation across groups is **heteroscedastic**.
   * Tests for **linearity** check for a linear relationship between two variables.
   * **Independence** describes a situation wherea change in the value of one variable does not affect the value of another variable. Non-independence arises from biased or non-random sampling.
2. Types of Data

* **Categorical Data:**
  + **Nominal –** discrete categories of data with no ranking or order – often expressed as a word. Nominal data is often summarized as percentages.
    - Examples: *Species Names, Thallus Morphology, Photobiont Type, Presence/Absence, etc.*
  + **Ordinal –** categories of data that are ranked or ordered from smallest to largest.
    - Examples: *Photobiont specificity (low/medium/high)*
* **Numerical Data:**
  + **Continuous** *–* **Scale Data** in JASP – measurement data that covers a wide, potentially infinite, range of values.
    - Examples: *Distribution, Density*
  + **Discrete** – count data that uses whole numbers or distinct values with the same distance between each integer.
    - Examples: *Number of species with apothecia*

1. Parametric Statistical Analyses – statistical tests that assume normal distribution, homoscedasticity, non-linearity, and independence.

* **Linear Regression/Correlation** *-* used to identify relationships between two or more continuous variables; to determine and compare the strength of relationships; or to calculate unknown values. In a linear regression the values of the independent variables are non-random. In correlation tests both continuous variables are randomly sampled from the population. As a reminder, correlation does NOT equal causation.
  + **Pearson’s r** – a parametric measure of correlation that assumes normality and linearity. This test is sensitive to outliers.
  + Null Hypotheses (H0): *There is no significant relationship between the response variable and the predictor variable – OR – slope 0*
  + Alternate Hypothesis (H1): *There is a significant relationship between the response and predictor variable – OR – slope = 0*

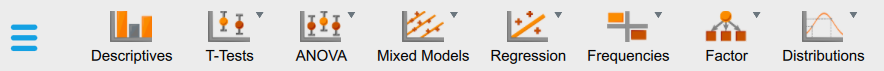
* **Independent Sample T-Test** - used to compare the means of a measurement variable of two groups of nominal data.
  + **Student’s t-test –** a two sample t-test used for data with equal variance in the two groups.
  + Null Hypothesis (H0): *There is no significant difference in the mean of the measurement variable for the two groups.*
  + Alternate Hypothesis (H1): *There is a significant difference in the mean of the measurement variable for the two groups.*
* **Analysis of Variance (ANOVA) –** used to compare the means of a measurement variable two or more groups of a nominal variable.
  + **One-way ANOVA –** one measurement variable and one nominal variable
  + **Two-way ANOVA –** one measurement variable and two nominal variables
  + Null Hypothesis (H0): *There is no significant difference in mean of the measurement variable for each group* - OR - *the mean for each group is the same.*
  + Alternate Hypothesis (H1): *There is a significant difference in the mean of the measurement variable for each group – OR – the mean for each group is not the same.*
* **Tukey Post Hoc Test –** used to identify pairs of means that are significantly different after rejecting the null hypothesis of an ANOVA.

1. Non-Parametric Statistical Analyses - statistical tests that may not assume normal distribution, homoscedasticity, non-linearity, and/or independence.

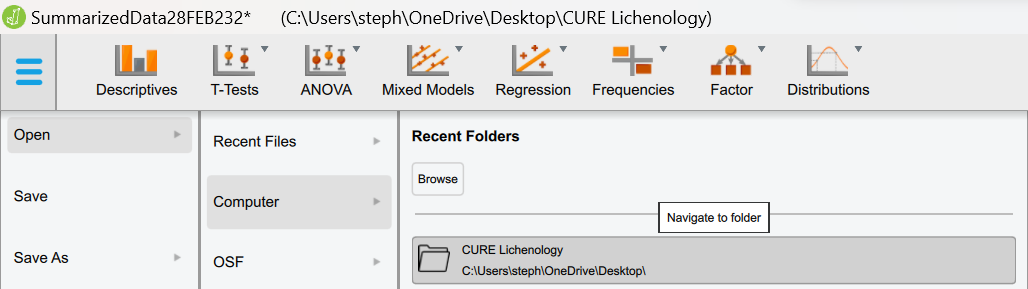
* **Correlation –** used to identify relationships between two or more continuous variables; to determine and compare the strength of relationships; or to calculate unknown values. In correlation tests both continuous variables are randomly sampled from the population. As a reminder, correlation does NOT equal causation.
  + **Spearman’s rho –** non-parametric measure of correlation used to test relationships between two ordinal variable or one ordinal variable and one continuous variable. This test *does not* assume normality or homoscedasticity. Spearman’s rho can used for ordinal data.
  + Null Hypotheses (H0): *There is no significant relationship between the response variable and the predictor variable – OR – slope 0*
  + Alternate Hypothesis (H1): *There is a significant relationship between the response and predictor variable – OR – slope = 0*
* **Independent Samples T-Test –** used to compare the means of a measurement variable of two groups of nominal data.
  + **Welch’s T-Test –** a two sample t-test used to compare the mean of independent groups of ordinal or continuous data that are heteroscedastic.
  + **Mann-Whitney T-Test –** a two sample t-test used to compare the median of independent groups of ordinal or continuous data that both are heteroscedastic and have a non-normal distribution.
  + Null Hypothesis (H0): *There is no significant difference in the mean/median of the measurement variable for the two groups.*
  + Alternate Hypothesis (H1): *There is a significant difference in the mean/median of the measurement variable for the two groups.*
* **Paired Samples T**-**Test –** used to compare one measurement variable among two groups of nominal data. This test *does not* assume homoscedasticity or that the measurement within each group has a normal distribution. This test should only be used when there is a single observation for each combination of nominal data. If this is not the case a **two-way ANOVA** should be used.
  + **Student’s Paired T-Test -** The student’s paired t-test *does* assume a normal distribution in the difference between pairs.
  + **Wilcoxon Signed-Rank Paired T-Test** – similar to a student’s paired t-test, this test used to compare one measurement variable among two groups of nominal data with extremely non*-*normal distribution in the difference between pairs.
  + Null Hypothesis (H0): *There is no difference in the mean of the paired observations.*
  + Alternate Hypothesis (H1): *There is a significant difference in the mean of the paired observations.*
* **Analysis of Variance (ANOVA) -** used to compare the means of a measurement variable two or more groups of a nominal variable.
  + **Kruskal-Wallis Test –** a **one-way or two-way ANOVA** but is insteadused to compare the means of one ordinal variable and one nominal variable with a non-normal distribution. This test should NOT be used if the standard deviation of your variables is heteroscedastic – instead use a **Welch’s ANOVA.**
  + **Welch’s ANOVA –** a one-way ANOVA used to compare the means of one ordinal variable and one nominal variable when the standard deviation of the variables is heteroscedastic.
  + Null Hypothesis (H0): *There is no significant difference in mean of the measurement variable for each group* - OR - *the mean for each group is the same.*
  + Alternate Hypothesis (H1): *There is a significant difference in the mean of the measurement variable for each group – OR – the mean for each group is not the same.*

**Part 2. Conducting Statistical Analyses in JASP**

1. Uploading your dataset to JASP
2. Open JASP. In the ribbon at the top of the program window you will see the **Main Menu** (three blue bars) along with all of the **Primary Modules** for various analyses.



1. To load your dataset, click the Main Menu 🡪 Open 🡪 Computer 🡪 Browse. The browse button will allow you to locate and select the .csv file you created during the Dataset Generation and Data Management tutorial.

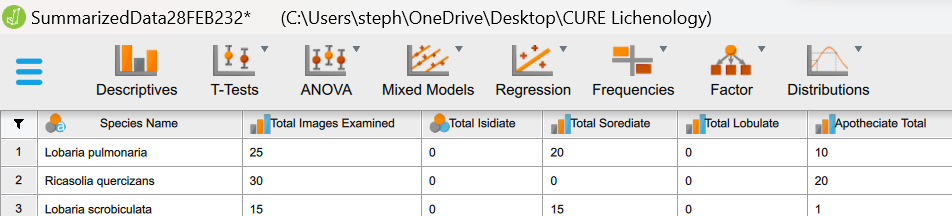


1. A screenshot of a computer

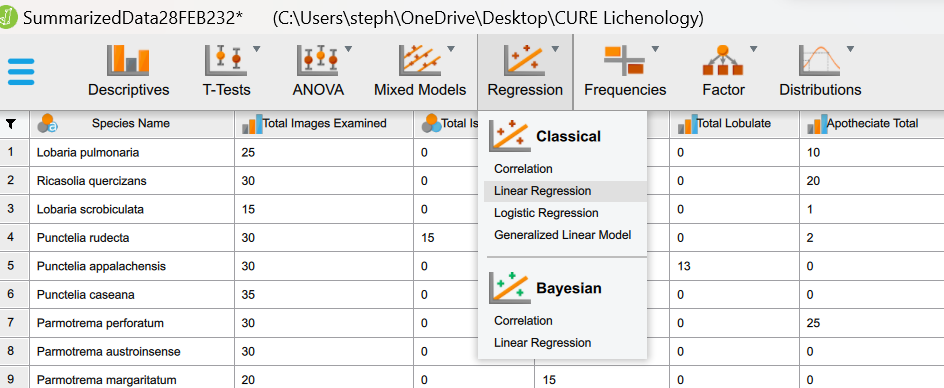
   Description automatically generatedOnce you have loaded your dataset your screen will look similar to the image below.

**It is important to note that ANY changes you make and save to the .csv file you uploaded to JASP in Excel will be automatically reflected in JASP.**

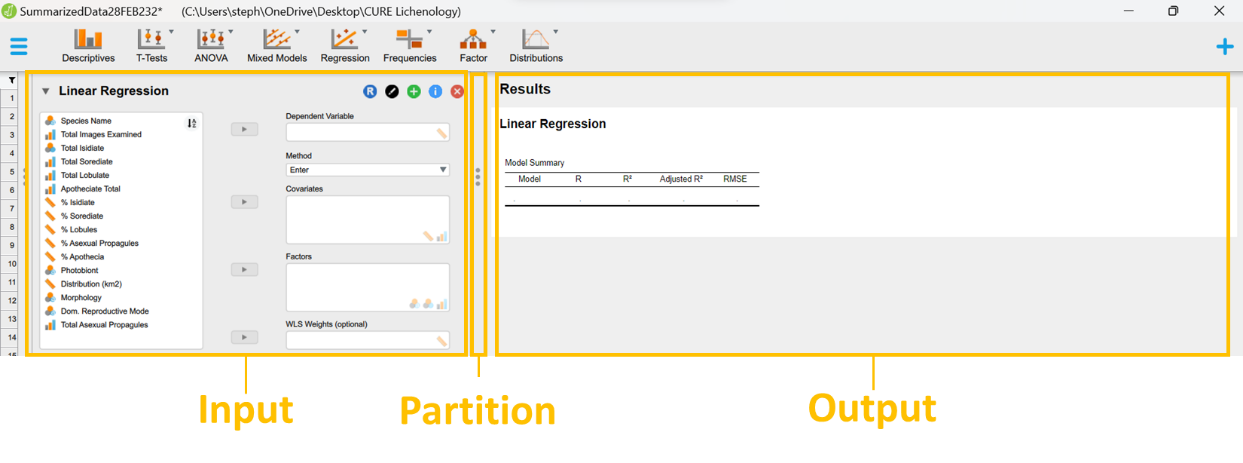
1. Double check that JASP has read your variables into the program using the appropriate data type. To change the data type for your column simply click the icon next to your column name and select the correct data type for your data.



1. Use the Main Menu (three blue bars) to save your JASP project now.
2. Linear Regression
3. Select the **Regression** module and the **Linear Regression** module under the dropdown menu.



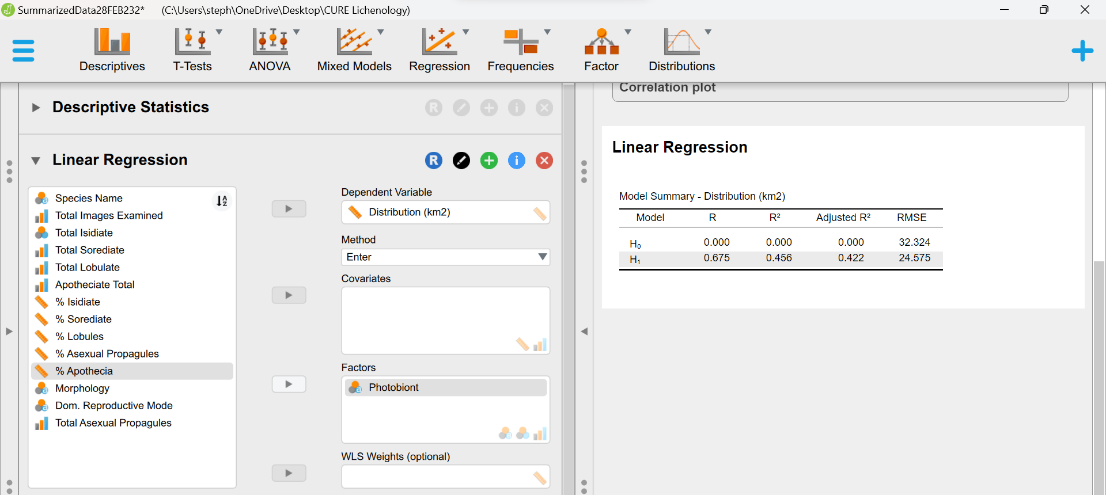
1. Once the Linear Regression module has been selected the JASP window will shift to become partitioned into two sections. On the left there is the **Input** with the variables from your dataset and the parameters that can be used to build a model. On the right there is the **Output** which will visualize the results from analyses and changes to the model in real time. To access your dataset or resize the input and output hold grab and drag the three circles on the **partition** accordingly.

****

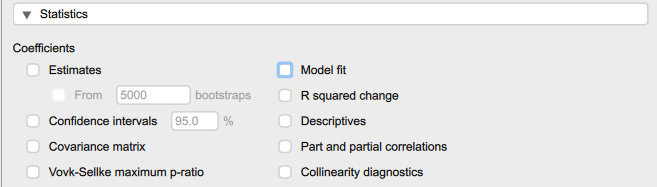
1. At the top of every analysis module, you will notice the following icons each with a different function.

|  |  |
| --- | --- |
|  | This dropdown arrow allows you to collapse the analysis module to free up space for other analysis modules. |
|  | The R icon displays the syntax (code) for the analysis that JASP is running using the statistical program language R. |
|  | The pen icon allows you to change or customize the name of the name of the statistical analysis you are conducting. |
|  | The plus sign icon allows you to create a copy of our analysis. |
|  | The information icon provides you with information on the analysis module including information on the assumptions, input, plots, statistics, output, and references. |
|  | The X icon closes the analysis module. |

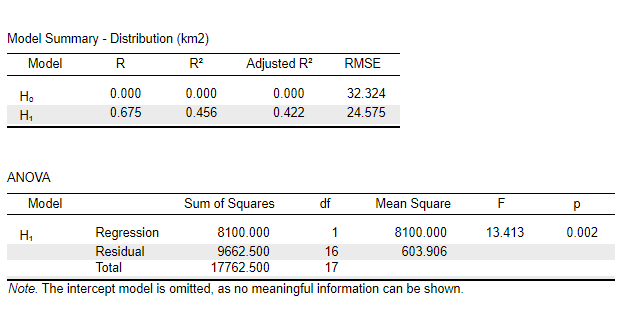
1. To build a linear regression model use the arrows or drag your dependent variable into the **Dependent Variable** box and your independent variable into the **Factors** variable box. In this example “Distribution (km2)” is the **dependent** variable and “Photobiont” is the **independent** variable.



1. Next check the **Model fit** box under the **Statistics** dropdown menu to run an ANOVA using the variables in the linear regression model.



1. JASP will automatically generate summary tables for the linear regression and ANOVA in the Output. The statistics we will need to consider are the **p-values**, **F**, **degrees of freedom**, and the **adjusted R2**.



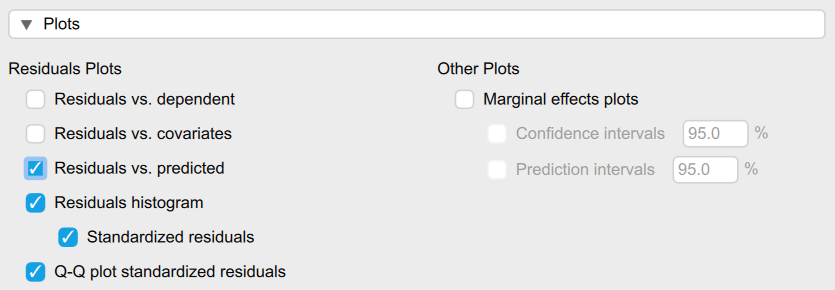
**p-value** - used to reject or accept the null hypothesis. It is the probability of getting the observed test statistic or one that is more extreme if the null hypothesis (H0) is true. In this model, if our alternate hypothesis obtains a p-value of < 0.005 then we can reject the null hypothesis.

**F-ratio –** a measure of the difference in the mean of two or more groups. It is calculate using the degrees of freedom (within-group and among-group variance) in the model. A F-ratio close to 1 indicates there is no difference in the means of the groups and we accept the null hypothesis.

**Degrees of Freedom (df)** - the number of independent variables in a model that can be changed without influencing the model. This statistic provides us with insight into the accuracy of our model. The regression df equate to *among-group variance* and the residual df equate to *within group variance*.

**Adjusted R2** – is the amount of variance in our dependent variable that can be explained by our independent variable(s). It is a measure of the strength of the associations in our model. The higher the R2 , the stronger our relationship is and vice versa. We only consider the adjusted R2 because if we add dependent variables to our model the adjusted R2 will only increase if those variables are meaningful to our model. In contrast, R2 will increase as we add dependent variables regardless of the significance of other variables to our model – this is referred to as overfitting.

1. To check that the assumptions for a linear regression are not being violated by the model examine the residual plots for this model. To do this use the **Plots** dropdown menu (found below the variable boxes) and check the following boxes listed under **Residuals Plots: Residuals vs predicted**, **Residuals Histogram**, and **QQ standardized residuals.**



**Residuals** are the difference (or deviation) in the observed values and estimated values of the dependent variable when compared to an estimated linear regression equation. Residuals can be used to estimate the error in a model and determine what assumptions may or may not violated in that model.

Chart, scatter chart

Description automatically generated**Residuals vs. Predicted** - this graph is used to visualize **homoscedasticity**. If the fit line in this graph displays any kind of slope it would indicate that the assumption of homogeneity of variance has been violated. In the example to the right, we have a fit line with no change in slope indicating that the assumption of homoscedasticity has not been violated.

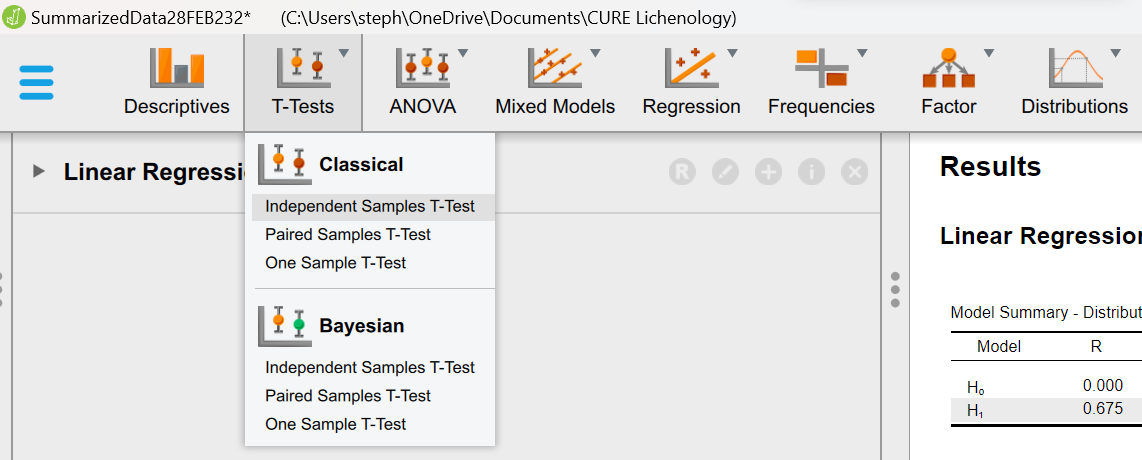
**QQ Plot/Residuals Histogram** - these graphs are used to visualize **normality**. If our residuals in the **histogram** produce a bell-shaped or Gaussian curve, that indicates the assumption of normality has not been violated. The example below isn’t perfect but close enough. In the **QQ plot** if our data fit to the line that indicates the assumption of normality has not been violated. It is not unusual for points at the far ends of the fit line to stray a bit. Again, the example below isn’t perfect, but not extreme enough to indicate the assumption of normality has been Chart, histogram

Description automatically generatedviolated.

1. Now that the model has been built and we have checked our assumptions we can report our results. Here is an example using the model that was created in this tutorial:

We recovered a significant relationship between photobiont type and species distribution (F(1, 16) = 13.41, p = 0.002). In fact, 44.2% of variation in distribution can be explained by photobiont type (R2 = 0.442).

1. Use the main menu to save your JASP file now.
2. Independent Samples T-Tests
   * + 1. Select the **T-Test** module followed by the **Independent Samples T-Test** module under **Classical**.

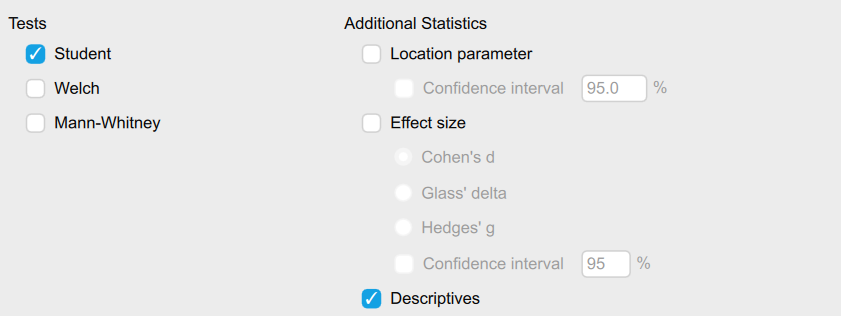


* + - 1. To build the model use the arrows or drag your dependent variable into the **Dependent Variable** box and your independent variable into the **Factors** variable box. In this example “Distribution (km2)” is the **dependent** variable and “Photobiont” is the **independent** variable.

Graphical user interface, text, application, chat or text message

Description automatically generated

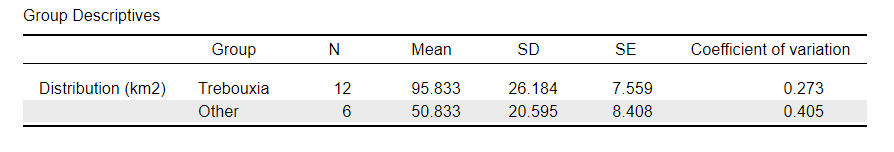
* + - 1. Next check the Descriptives box under Additional Statistics which can be found below the Input variable boxes.



* + - 1. JASP will automatically update the output with summary tables for t-test and the descriptive statistics.

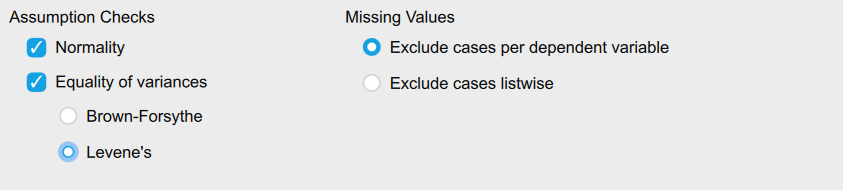


The statistic to consider is the **p-value.** In this model the p-value is 0.002 which is < 0.005 indicating that we can reject the null hypothesis.

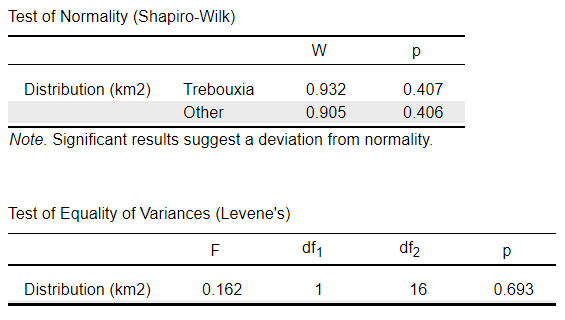


Later, report the **t-value** (ratio of difference between the sample mean and the population mean), the **degrees of freedom (df)**, and **p-value** from the Independent Samples T-test summary table, in addition to the **mean** and **standard deviation (SD)** from the Group Descriptives summary table.

* + - 1. To check that the assumptions for a t-test have not been violated use the **Assumptions Checks** located below the input variable boxes. Check both the **Normality** and **Equality of Variance** boxes and select **Levene’s** under Equality of variances.



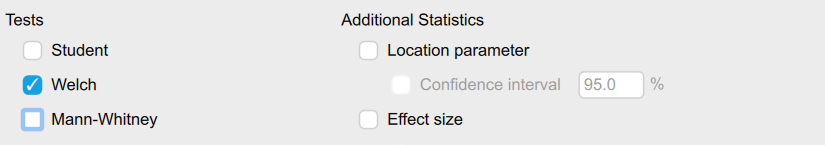
* + - 1. JASP will automatically update the output with the results from those assumption checks.



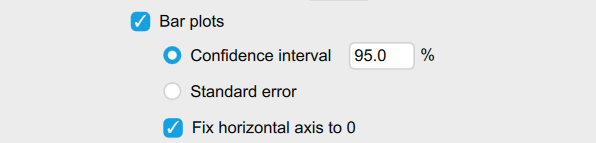
**Shapiro-Wilk Test** – uses the mean of the dataset to test for **normality**. The **null hypothesis (H0)** for a Shapiro-Wilk test is that the values for the variables used in the model are normally distributed population. The **alternative hypothesis (H1)** for the test is that the variables used in the model are not normally distributed. We can use the **p-value** of the Shapiro-Wilk test to accept the null hypothesis (p-value > 0.05) or reject the null hypothesis (p-value < 0.05). In the example above the p-values for the Shapiro-Wilk test of the variable “Distribution (km2)” across both photobiont types (*Trebouxia* and Other) are not significant (p-value > 0.005). Therefore, we can accept the null hypothesis that the values for the dependent variable “Distribution (km2)” are normally distributed among both groups (*Trebouxia* and Other) of the independent variable “Photobiont”. The assumption of normality has not been violated.

**Levene’s Test** – tests for equal variance (**homoscedasticity**) across all of the samples for dependent variable used in the model. The **null hypothesis (H0)** for a Levene’s test is that variance is equal across all of the samples. The **alternative hypothesis (H1)** for the test is that variance is not equal across all of the sample. **T**he **p-value** of the Levene’s test can be used to accept the null hypothesis (p-value > 0.05) or reject the null hypothesis (p-value < 0.05). In the example above the p-value the test of the variable “Distribution (km2)” is not significant (p-value > 0.005). Therefore, we can accept the null hypothesis that there is homoscedasticity across all samples of the dependent variable “Distribution (km2)”. The assumption of homoscedasticity has not been violated.

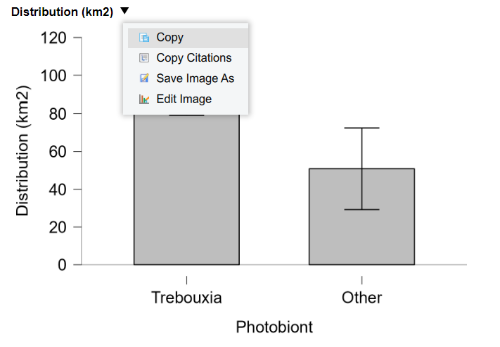
* + - 1. If your data **DOES NOT** violate the assumptions of normality and homoscedasticity you can move to Step 8. If your data **DOES** violate the assumptions of normality and/or homoscedasticity you will have to conduct a non-parametric t-test. A **Welch’s t-test** can be used if your data violate the assumption of heteroscedasticity. A **Mann-Whitney t-test** if your data violate the assumption of normality and heteroscedasticity. Either test can be select by checking the **Welch** or **Mann-Whitney** boxes under **Tests** below the input variable boxes.



* + - 1. The results from a t-test can be visualized using a **bar plot**. To generate a bar plot of our model, click the **Bar plot** box under **Additional Statistics** below the input variable boxes.



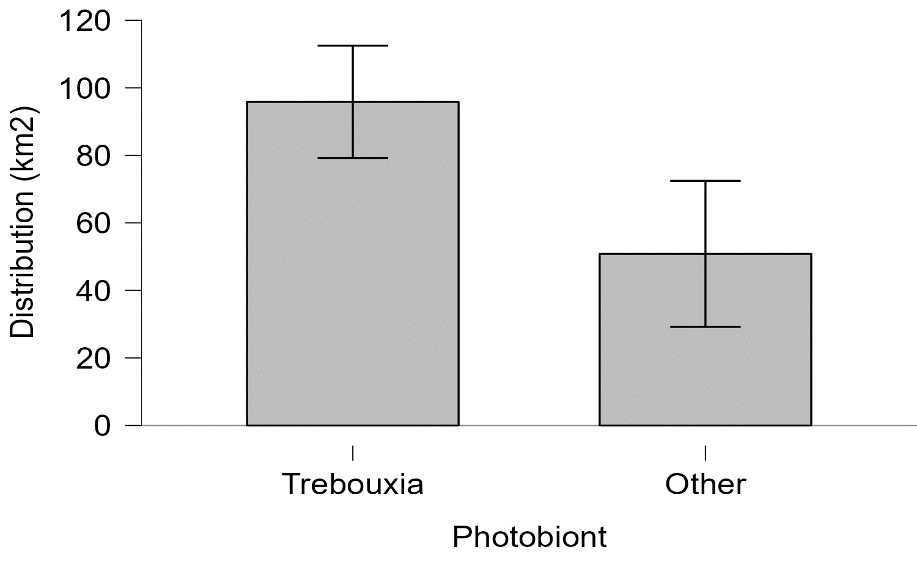
* + - 1. JASP will automatically generate a bar plot in the output. To copy, save, or edit the bar plot use the dropdown menu next to the title of your bar plot.



* + - 1. Now that the t-test has been performed, the assumptions have been checked, and a graph has been made you are ready to report your results. Below is an example report of t-test results using the model that was created in this tutorial.

We recovered a significant difference in the distribution between lichen species with photobionts in the genus *Trebouxia* (Mean = 95.833, SD = 26.184) and those with photobionts in other genera (Mean = 50.833, SD = 20.595); t(16) = 3.662, p-value = 0.002.

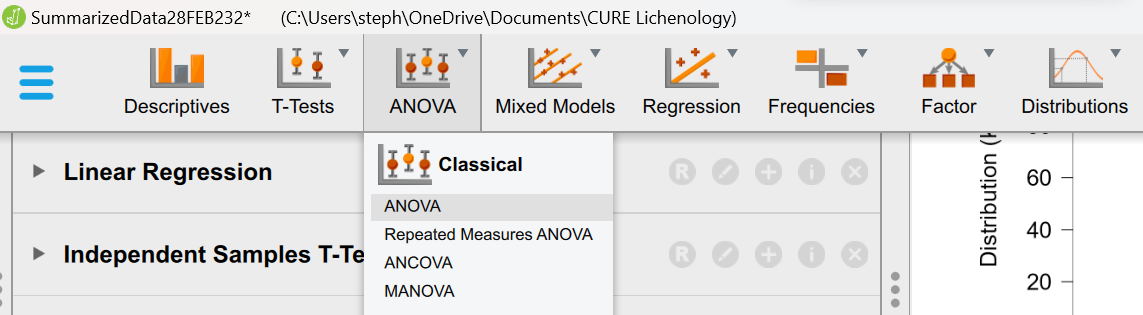
The graph from this test can be presented in your report as a figure. Be sure to give your figure an appropriate caption. Figures and captions should be comprehensive enough to stand alone. To the right is an example using the graph for the model the model that was created in this tutorial.



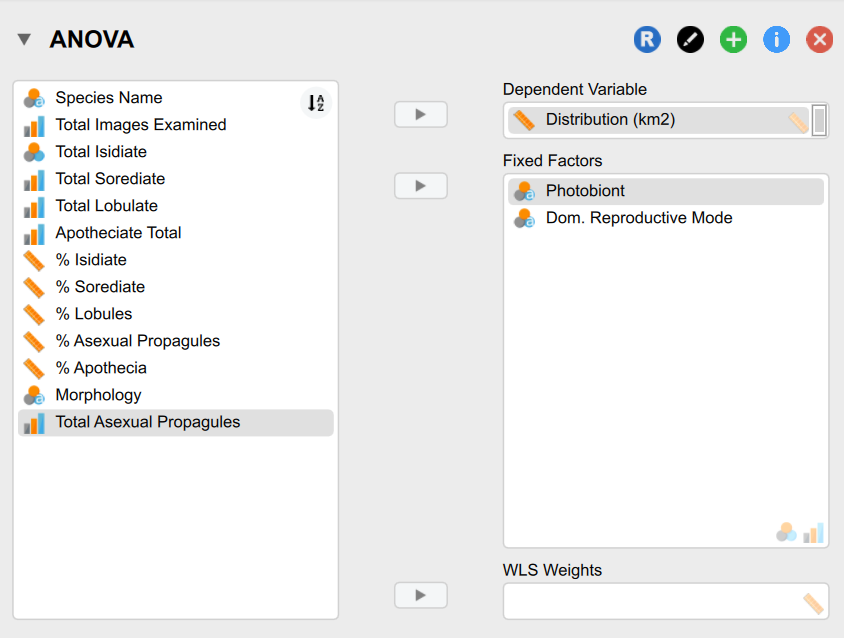
**Figure 1.** Bar plot displaying the mean distribution (km2) of lichen species with photobionts in the genus Trebouxia and species with photobionts in other genera that include Asterochloris/Symbiochloris. A t-test was performed and it was found that there is a significant difference in the distribution between lichen species with photobionts in the genus Trebouxia and those with photobionts in the genera Asterochloris and Symbiochloris; t(16) = 3.662, p-value = 0.002. Error bars were calculated using a 95% confidence interval.

* + - 1. Use the main menu to save your JASP file now.

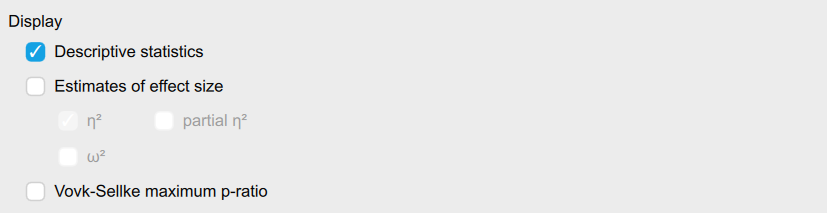
1. ANOVA
2. Select the **ANOVA** module followed by the **ANOVA** module under Classical.



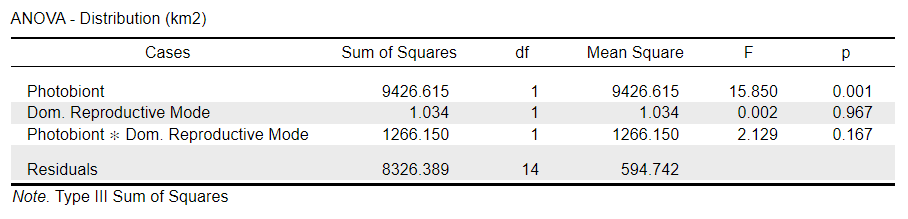
1. To build our model use the arrows or drag your dependent variable into the Dependent Variable box and drag your independent variables into the Fixed Factors box. In this example “Distribution (km2)” is the dependent variable and “Photobiont” and “Dominant Reproductive Mode” are the independent variables. By incorporating two categorical variables into our model, we are performing a **two-way ANOVA.**



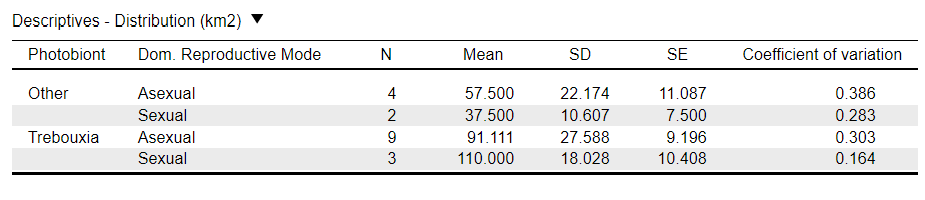
1. Next check the **Descriptive statistics** box under **Display** which can be found below the input variable boxes.



1. JASP will automatically update the output with summary tables for the ANOVA and the descriptive statistics.

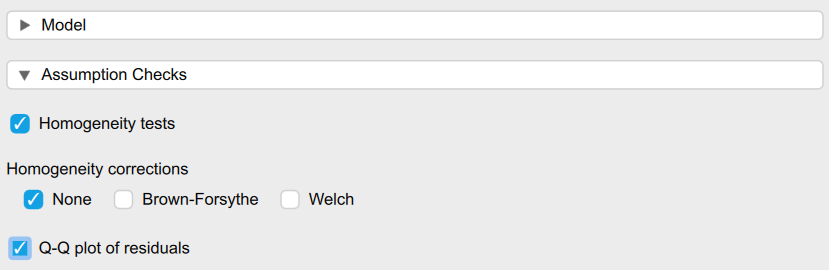


The statistic we need to consider in an ANOVA is the **p-value.** In this model we have 3 p-values. The first p-value (0.001) reflects the significance of the effect of photobiont on distribution. The second p-value (0.967) reflects the significance of the effect of dominant reproductive structure on distribution. And the third p-value (0.167) reflects the significance of the effect of the interaction between photobiont and dominant reproductive structure on distribution. Only the p-value for the test of the effect of photobiont on distribution (p-value = 0.001) was < 0.005 indicating that for we can only **reject** the null hypothesis for that relationship. For the remaining variables in this model we **accept** the null hypothesis (p-value = > 0.005). Later you will report the results of our model using the **p-value** along with the **F-ratio** and the **degrees of freedom (df)**.

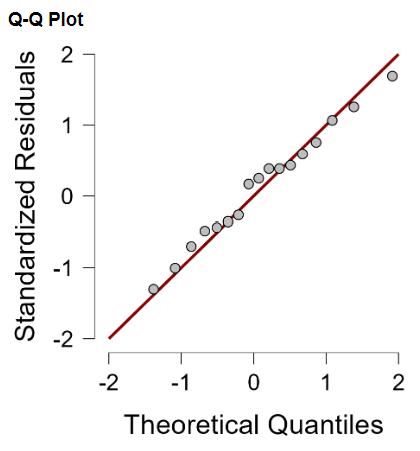
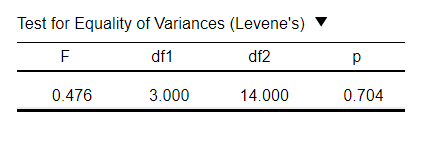


The Descriptives summary table can be presented in a report as table or part of the supplementary data.

1. To check that the assumptions for an ANOVA have not been violated we can use the **Assumptions Checks** dropdown menu located below the input variable boxes and the Model dropdown menu. Check both **Homogeneity tests** and the **QQ plot of residuals** boxes.

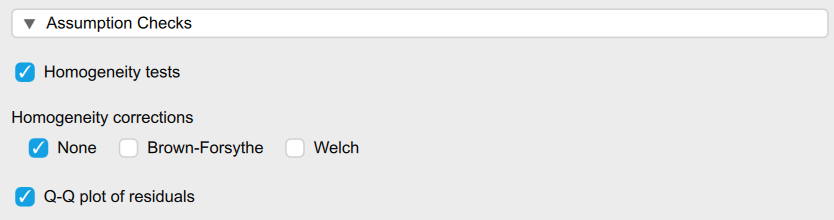


1. JASP will automatically update the output with a summary table for a Levene’s test (left) and a QQ Plot (right). Based on the p-value of the Levene’s test (0.0704 > 0.005) and the fit of the model residuals to the line in the QQ plot the assumptions of homoscedasticity and normality have not been violated.

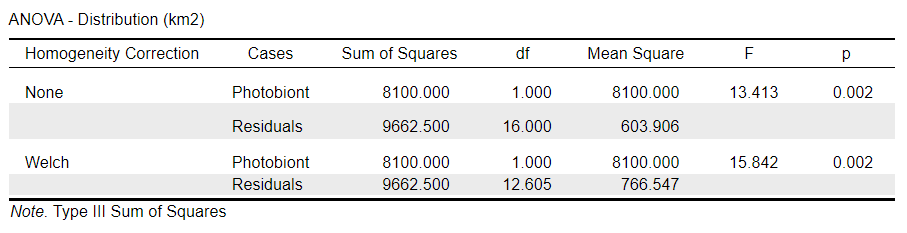


1. If your model DOES NOT violate the assumptions of homoscedasticity and normality continue on to **Step 8**. If your model DOES violate the assumptions of homoscedasticity and normality you will have to perform a non-parametric ANOVA.

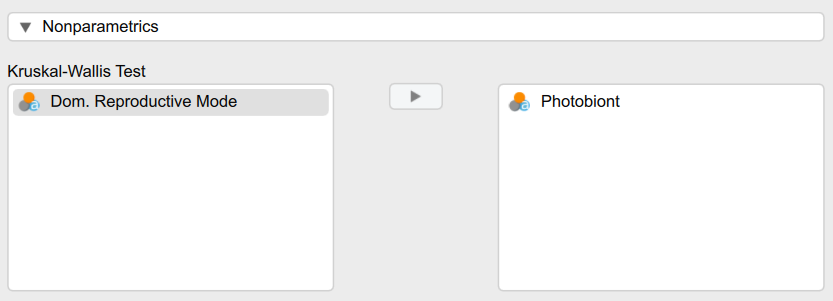
If you are performing a **one-way ANOVA** and your model violated only the assumption of homoscedasticity perform a **Welch’s t-test**. To perform a Welch’s t-test check the **Welch** box under **Homogeneity corrections** in the **Assumption Checks** dropdown menu to perform a Welch’s t-test.



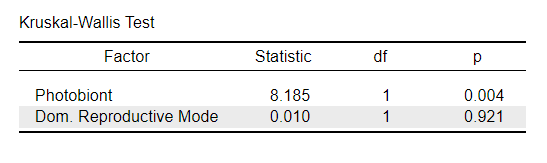
JASP will automatically update the ANOVA summary table in the Output displaying the results for the corrected model.



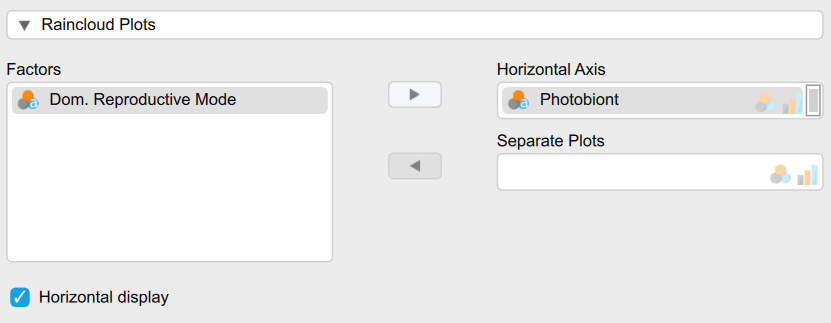
If you are performing a **one-way or two-way ANOVA** and your data is homoscedastic but has violated the assumption of normality perform a **Kruskal-Wallis Test.** To perform a Kruskal-Wallis Test use the **Nonparametrics** dropdown menu at the bottom of the Input.The independent variables you want to include in your model can be moved with the arrow buttons or dragged into the variables box on the right.



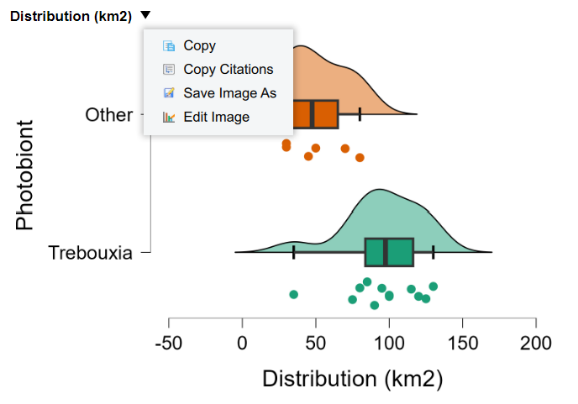
JASP will automatically generate a new Kruskal-Wallis summary table in the Output.



1. The results from an ANOVA can be visualized using a **Raincloud Plot**. Use **Raincloud Plots** dropdown menu to reveal the variable boxes. Next drag or use the arrow to move the **factor** (variable) you would like to plot into the **Horizontal Axis** box. To display horizontally check the **Horizontal display** box.



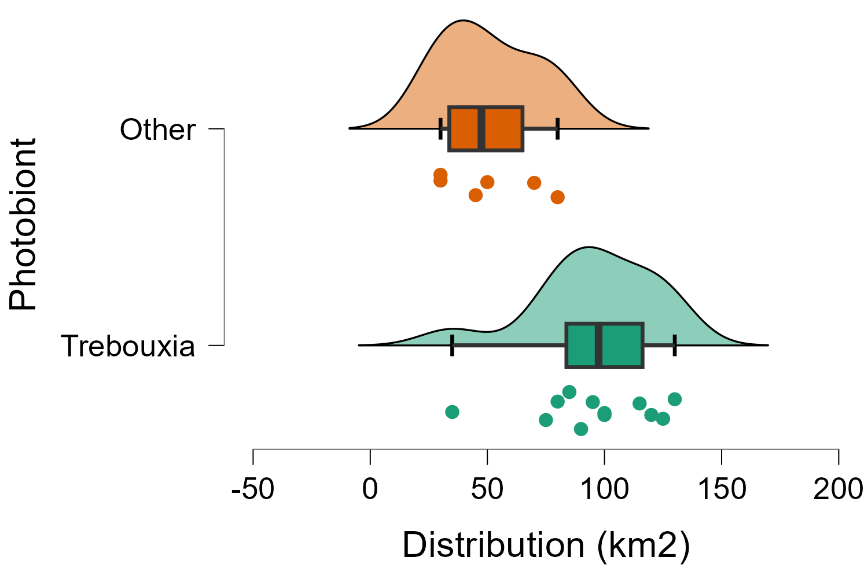
1. JASP will automatically generate a raincloud plot in the output. To copy, save, or edit the bar plot use the dropdown menu next to the title of your bar plot.



1. Now that the ANOVA has been performed and the assumptions have been met you are ready to report the results. Below is an example report of ANOVA results using the model created in this tutorial.

We found that photobiont type had a significant effect on species distribution size (F(1, 14) = , p = 0.002). In contrast, both dominant reproductive mode (F(1, 14) = 0.002, p = 0.967) and the interaction of photobiont type and dominant reproductive mode (F(1, 14) = 2.129, p = 0.167) had no significant effect on species distribution.

The graph from this test can be presented in your report as a figure. Be sure to give your figure an appropriate caption. **Figures and captions should be comprehensive enough to stand alone**. To the right is an example using the graph for the model created in this tutorial.



**Figure 2**. Raincloud plot displaying the distribution (km2) of lichen species with photobionts in the genus Trebouxia and species with photobionts in other genera that include Asterochloris/Symbiochloris. An ANOVA was performed, and it was found that photobiont type had a significant effect on species distribution (F(1, 14) = , p = 0.002).

1. Use the main menu (three blue bars) to save your JASP project now.

**Part 3. Reporting JASP in Publications**

Great work! You are now ready to include your findings in your research manuscript. To report JASP in the methods sections of your research manuscript you will need to describe the statistical tests you performed and note that JASP was the program used to perform those tests. When referencing JASP in your manuscript it is necessary to include which version of JASP you are using along with a citation for the publication that first introduced and described JASP.

Version Information

Version information for JASP can found under the **Main Menu** by clicking on **About**. JASP will automatically open a new window that will provide you with the version information.

Text

Description automatically generated

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., ... & Wagenmakers, E. J. (2019). JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software*, *88*, 1-17.

Reference for JASP

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., ... & Wagenmakers, E. J. (2019). JASP: Graphical statistical software for common statistical designs. Journal of Statistical Software, 88, 1-17.

Linear Regression

Here is an example for reporting a linear regression using the model created in this tutorial:

A linear regression was performed using JASP version 0.17.1 to test the relationship between photobiont type and distribution (km2) (Love *et al* 2019).

T-Test

Here is an example for reporting a t-test using the model created in this tutorial:

A two-sample t-test was performed using JASP version 0.17.1 to compare distribution (km2) in lichen species with green algal photobionts in the genus *Trebouxia* and species with photobionts in other genera that included *Asterochloris* and *Symbiochloris* (Love *et al* 2019).

ANOVA

Here is an example for reporting an ANOVA using the model created in this tutorial:

A two-way analysis of variance (ANOVA) was performed using JASP version 0.17.1 to investigate the effect of photobiont type and dominant reproductive mode (sexual vs. asexual) on the distribution (km2) of lichen species (Love *et al* 2019).

**References**

1. McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland.
2. Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., ... & Wagenmakers, E. J. (2019). JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software*, *88*, 1-17.
3. R Core Team (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.