# **Introduction:**

# The chi-square test for association tests for whether two categorical variables are associated. Another way to phrase this is that this test determines whether two variables are statistically independent. For this reason, this test is also often referred to as the chi-square test of independence. More specifically, it tests for the association/independence between two nominal/dichotomous variables. You can test for ordinal variables, but you will lose the extra information provided by knowing the order of the categories. This test does not distinguish between dependent and independent variables, although your study design might do so.

# For example, you could use a chi-square test for association to determine whether there is an association between whether a person exercises and the presence of heart disease (i.e., your two nominal variables would be "exercise", which has two groups – "exercises" and "does not exercise" – and "presence of heart disease, which also has two groups: "yes" and "no". If there is an association (positive or negative), you can also determine the strength/magnitude of this association. As another example, you could use a chi-square test for association to determine whether there is an association between brand preference and gender in terms of sports cars (i.e., your two nominal variables would be "car brand preferences", which has five groups – Audi, BMW, Land Rover, Mercedes and Porsche – and gender, which has two groups: "males" and "females". Again, if there is an association (positive or negative), you can also determine the strength/magnitude of this association.

**What does the chi-square test do?**

The chi-square test for association determines whether there is an association between two nominal variables. It does this by comparing the observed frequencies in the cells to the frequencies you would expect if there was no association between the two nominal variables. As the expected frequencies are predicated on there being no association, the greater the association between the two nominal variables, the greater you would expect the observed frequencies to differ to the expected frequencies. The converse is also true. The less the two nominal variables are associated, the closer the observed frequencies will be to the expected frequencies. Indeed, this is how the chi-square test for association works. It produces a statistic based on the overall "amount" of difference between the expected and observed frequencies. The further the observed frequencies are to the expected frequencies, the larger the test statistic, the greater the association and the more likely a statistically significant result (i.e., indicating that an association exists).

# **Assumptions of the Chi-Square:**

* Assumption #1: You have two categorical variables. A categorical variable can be either nominal or ordinal; however, if you have an ordinal variable it is better to use a different test statistic. Examples of nominal variables include gender (two groups: males or females), ethnicity (e.g., three groups: Caucasian, African American and Hispanic), profession (e.g., five groups: surgeon, doctor, nurse, dentist, therapist), and so forth.
* Assumption #2: You should have independence of observations. This means that there is no relationship between the observations in the groups of the categorical variables or between the groups themselves. Indeed, an important distinction is made in statistics when comparing values from either different individuals or from the same individuals. Independent groups (in a chi-square test for association) are groups where there is no relationship between the participants in any of the groups. Most often, this occurs simply by having different participants in each group.
  + For example, imagine that a teacher wants to know if there is an association between gender and a person's preferred learning medium. In such an example, you have two nominal variables: "gender" (with two groups: "males" and "females") and "preferred learning medium" (with two groups: "online" and "books"). Independence of observation means that no person in the female group can also be in the male group (and vice versa). Similarly, a person cannot prefer the online medium and books. They can only prefer either the online medium or books. This will be true of any independent groups. In actual fact, the 'no relationship' part extends a little further and requires that participants in both groups are considered unrelated, not just different people; for example, participants might be considered related if they are husband and wife, or twins. Furthermore, participants in Group A cannot influence any of the participants in Group B, and vice versa.
* Assumption #3: All cells should have expected counts greater than five.

## **Null and Alternative Hypotheses:**

The null hypothesis for this test is as follows:

H0: Observed Values = Expected Values; There is no relationship between the variables.

And the alternative hypothesis is:

HA: Observed Values ≠ Expected Values; There is a relationship between Variable X and Variable Y.

## **Example:**

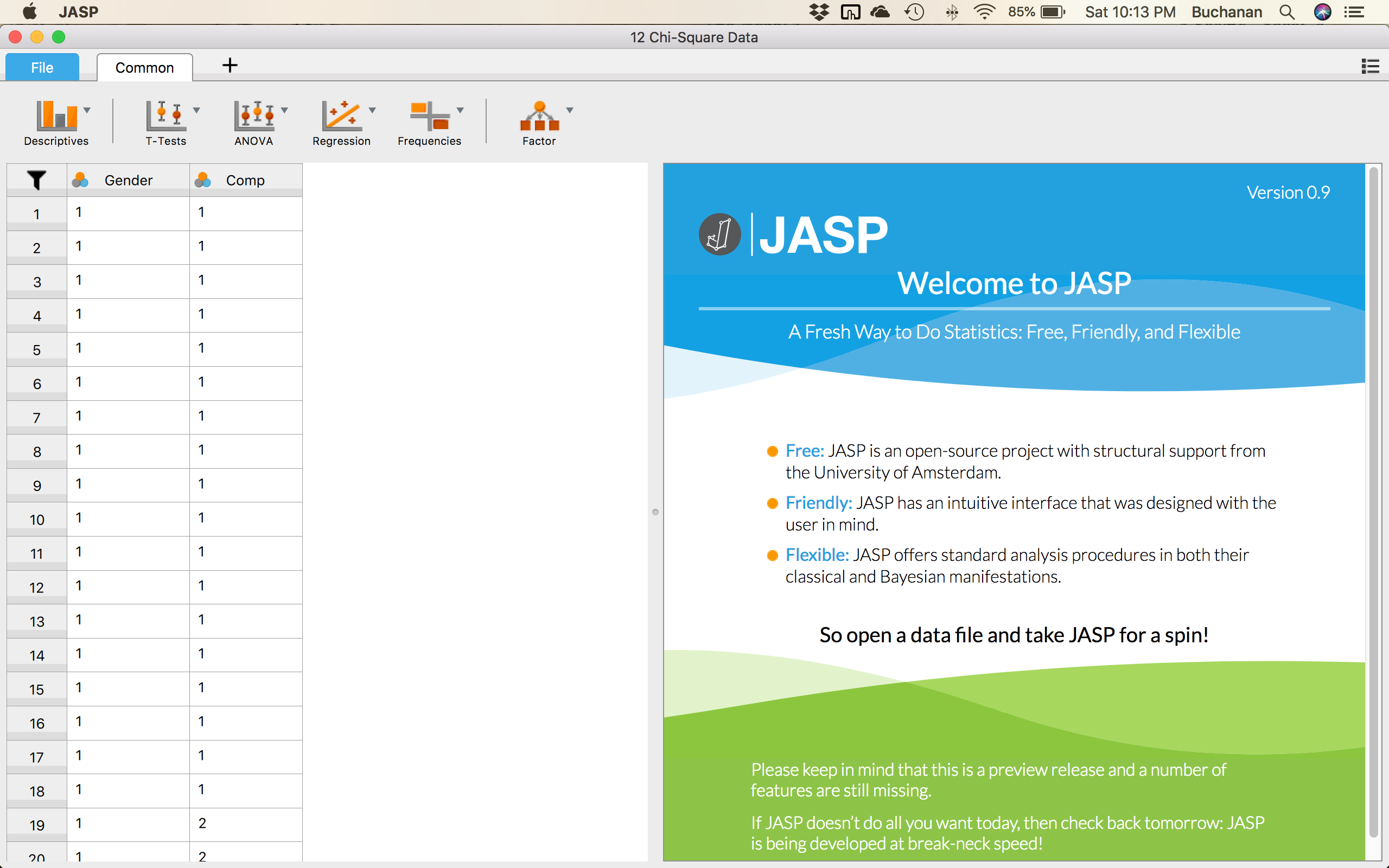
A researcher knows that in the general population of active individuals, males tend to engage in competitive sports whilst females prefer non-competitive sport/exercise. The researcher would like to investigate whether this is the case for males and females that are currently enrolled in an Exercise Science degree course. They asked 25 males and 25 females whether they predominately participate in competitive sport or non-competitive sport/exercise.

Whether participants predominantly participated in competitive or non-competitive sport was recorded in the variable, comp, whilst their gender was recorded in the variable, gender. In variable terms, the researcher wants to know what the association is between gender and comp.

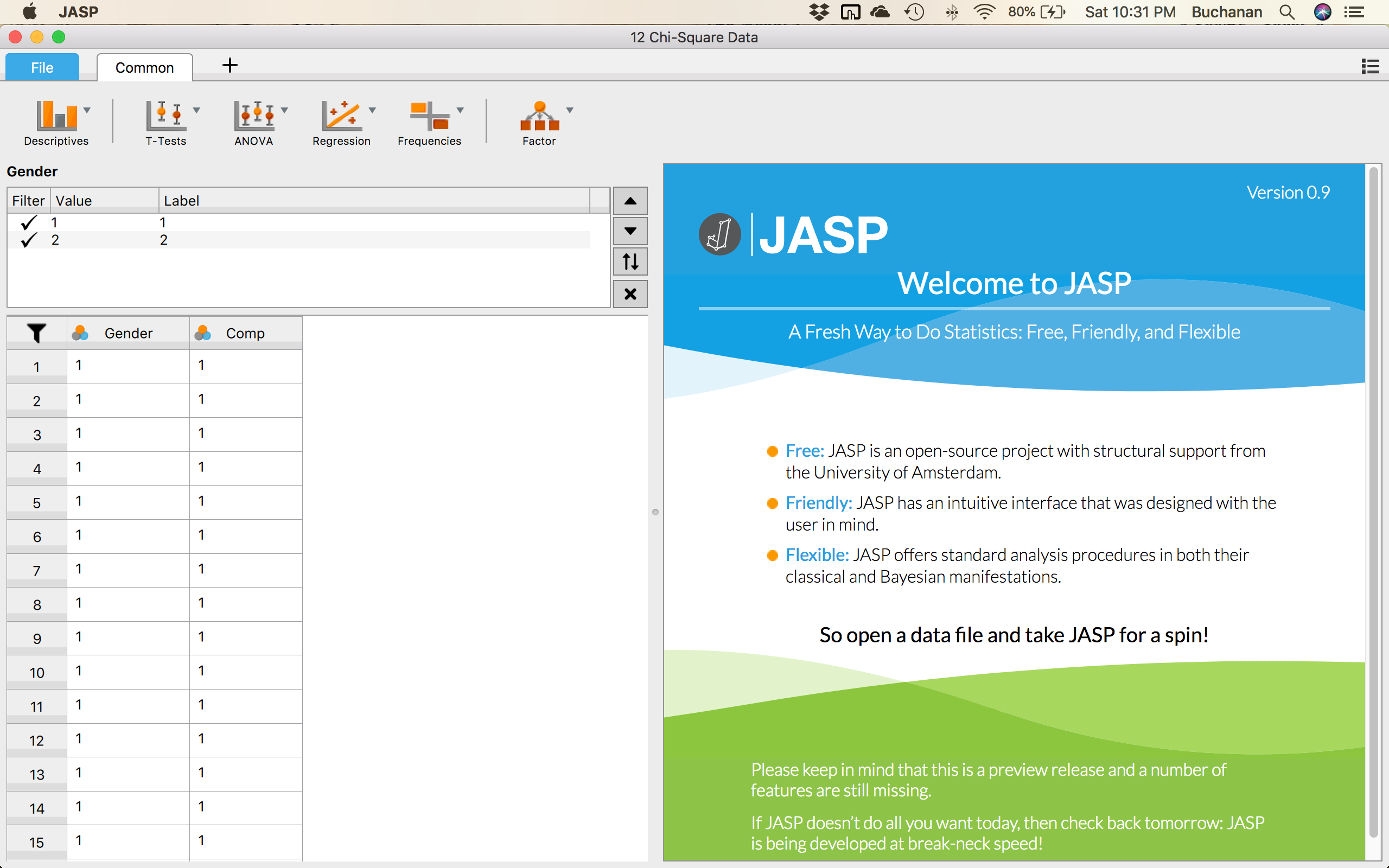
The data is set up such that each person gets their own row, which means that we have a gender and comp variable coded for each person indicating which category they below to for both variables.

To get started, open the dataset for this example in JASP. Remember, you can always use the previous help guides for greater detail in case you do not remember how to do something.

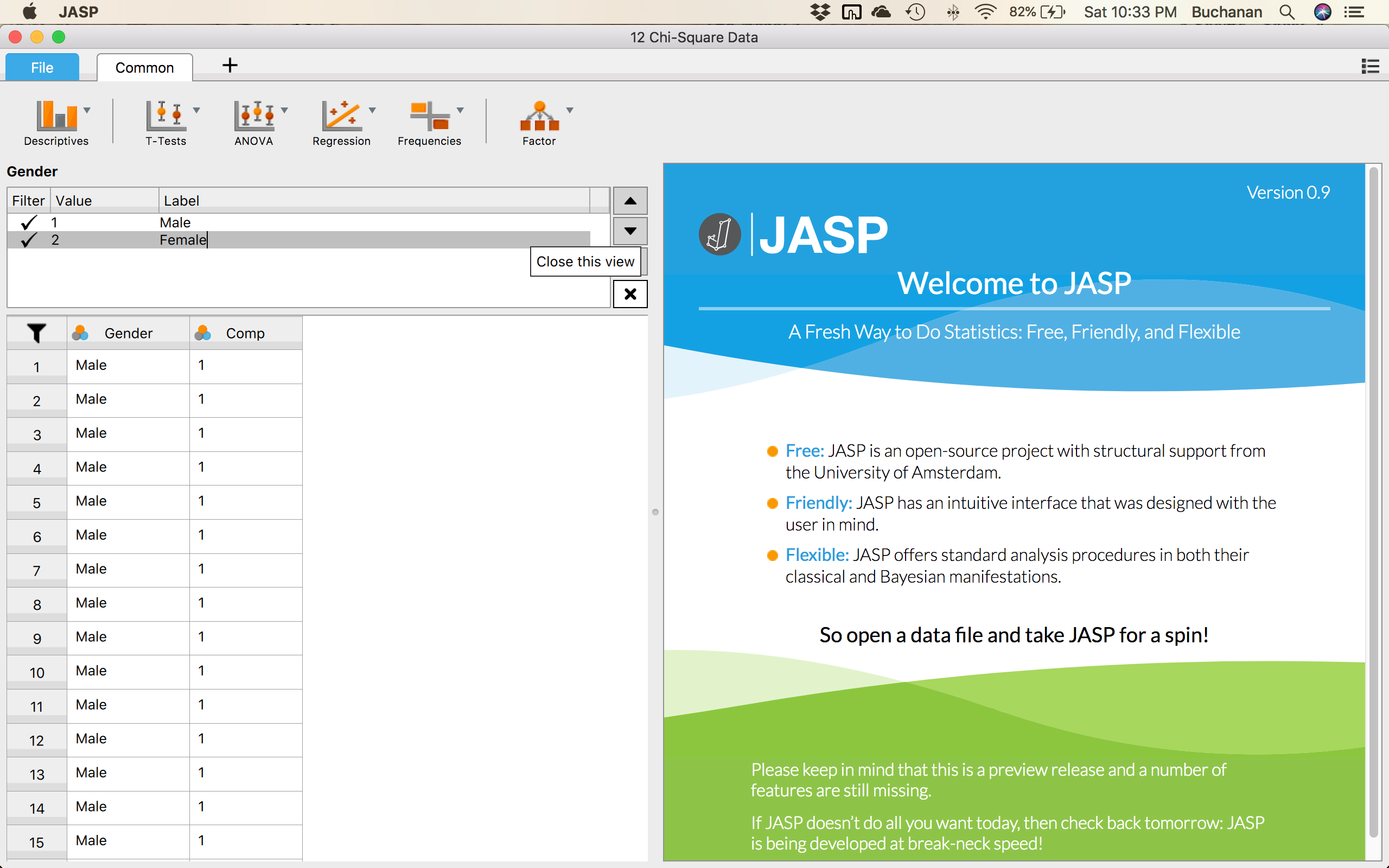
File 🡪 Open 🡪 Computer 🡪 Browse 🡪 Pick the Chi-Square Data.



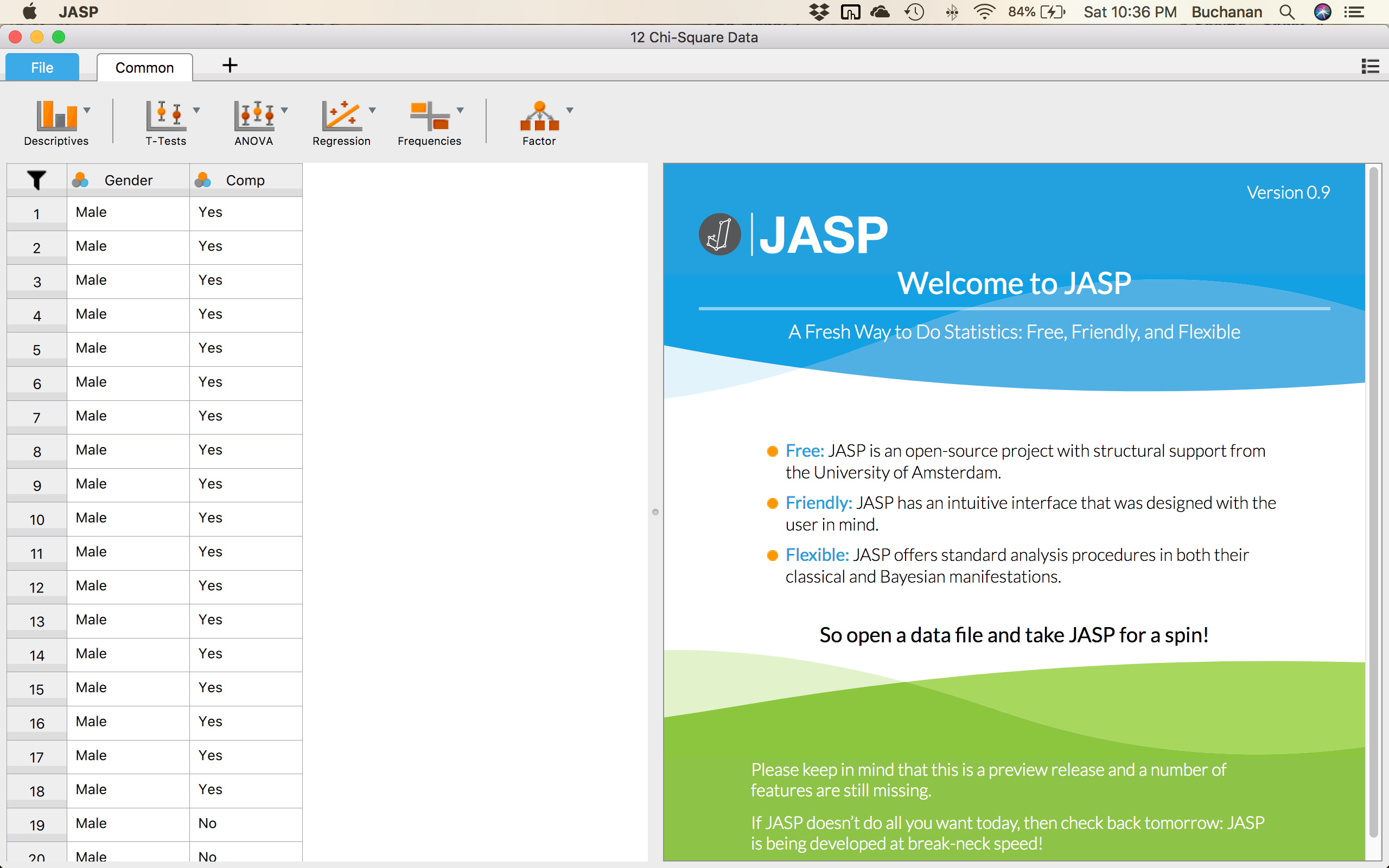
## Before we get started, we can fix our columns so that they include the labels. At the moment, it says 1 and 2. We could go into Excel and rename all of those. Or we can use the JASP option to give them labels. If you hover over gender, you will see a note saying “click here to change labels”. Click on it.



Change the 1 to Male and the 2 to Female then click the X to close the label window.



Do the same thing for the Comp where 1 is Yes and 2 is No.



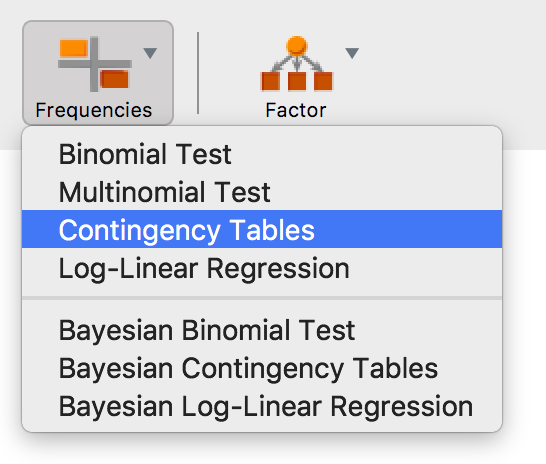
## **Check your assumptions:**

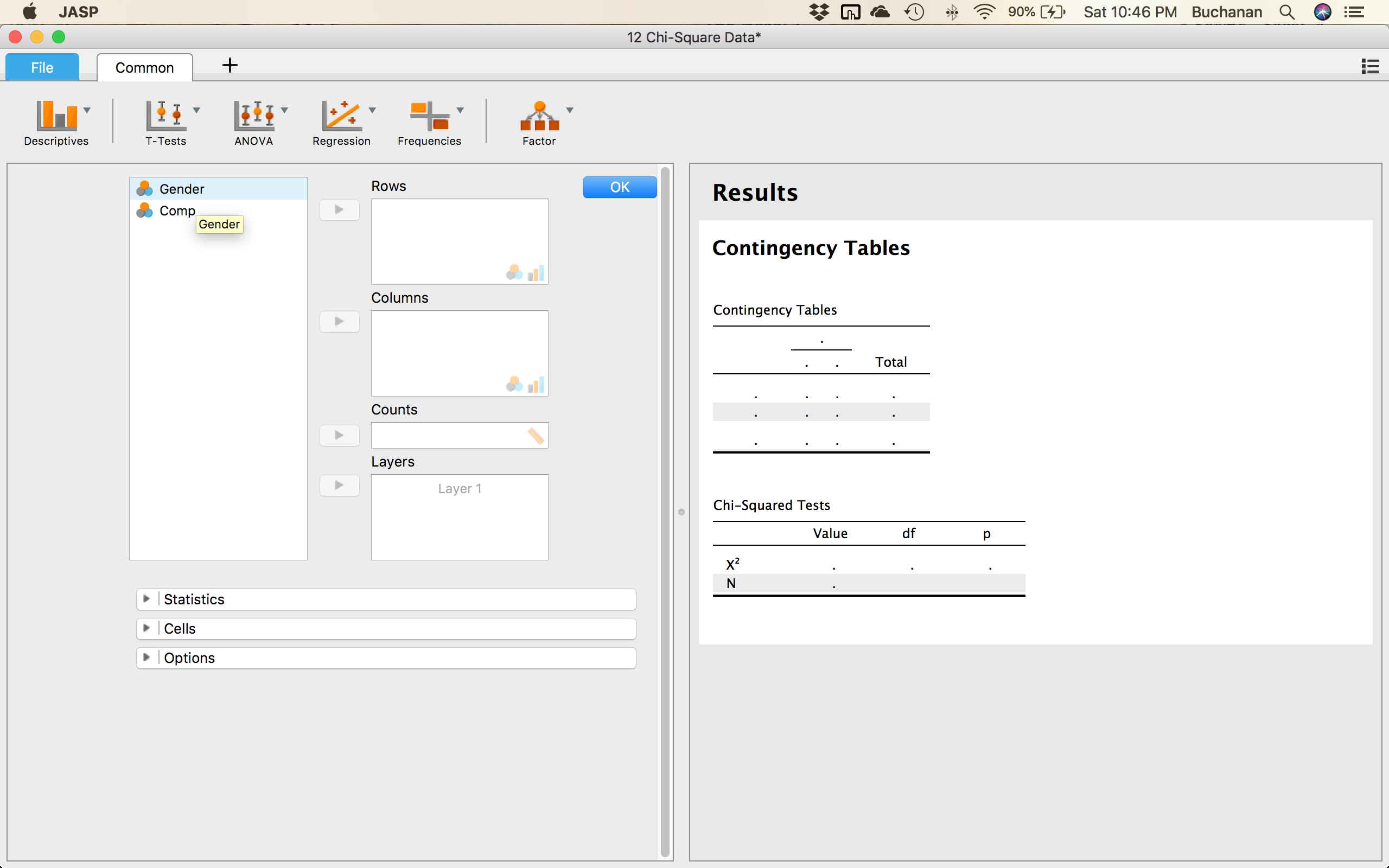
**Is the dependent variable categorical?**

Yes, we are using categorical style data.

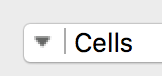
**Are there at least five participants expected in each cell?**

To get a table of the observed and expected frequencies of each group, you can click on  🡪 Contingency Tables.





Move gender and comp over one of them into the “Rows” box and the other into the “Columns” box. It does not matter which way you do it.

Click on Cells:  🡪 then check the Expected box  under “Counts”

| **Contingency Tables** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | **Comp** | | | |  | |
| **Gender** | |  | | **Yes** | | **No** | | **Total** | |
| Male |  | Count |  | 18.00 |  | 7.00 |  | 25.00 |  |
| Expected count |  | 14.00 |  | 11.00 |  | 25.00 |  |
| Female |  | Count |  | 10.00 |  | 15.00 |  | 25.00 |  |
| Expected count |  | 14.00 |  | 11.00 |  | 25.00 |  |
| Total |  | Count |  | 28.00 |  | 22.00 |  | 50.00 |  |
| Expected count |  | 28.00 |  | 22.00 |  | 50.00 |  |
|  | | | | | | | | | |

In this example, you can see that the expected count is greater than five in each of the four cells. The expected counts for males and females who predominantly participate in competitive sport are **14**, whilst the expected counts for males and females who predominantly participate in non-competitive sport are **11**. Therefore, the assumption that all cells should have expected counts greater than five has been met.

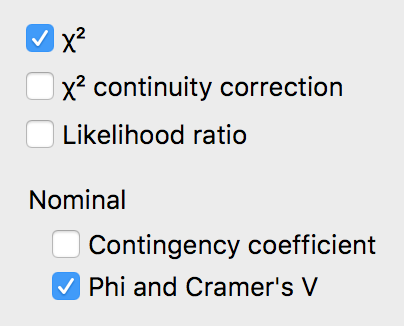
You could report the results for this assumption as follows: A chi-square test for association was conducted between gender and preference for performing competitive sport. All expected cell frequencies were greater than five.

From these results, you can see that for "males", the observed frequency (listed as count) was somewhat greater than expected count for "yes" to competitive sports, and lower than expected for "no" to competitive sports, and in "females", the other way around. This might lead you to suspect that there is an association between these two variables. You can test for this formerly in the next section.

## **The chi-square test:**

To get the chi-square test and all the other pieces that you’ll need, click .

The chi-square test is already selected, so also select Phi and Cramer’s V for effect size.



**Effect Size:**

| **Nominal** | | | |
| --- | --- | --- | --- |
|  | | **Value** | |
| Phi-coefficient |  | 0.322 |  |
| Cramer's V |  | 0.322 |  |
|  | | | |

Phi (φ) and Cramer's V are both measures of the strength of association of a nominal by nominal relationship. Phi is only suitable when you have two dichotomous variables. Phi and Cramer's V will provide the same answer when you have a 2 x 2 crosstabulation, although Phi is usually reported in these situations. Phi is not suitable for anything other than 2 x 2 tables, so in all other cases you should use Cramer's V. Both these measures can be interpreted in the same manner as a correlation (Phi ranges from -1 to +1). The major problem with these measures is that, under certain conditions, the maximum ranges can differ from -1 to +1. Here we could say that V = .32, which is similar to saying that the correlation or relationship between the variables is .32.

| **Chi-Squared Tests** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Value** | | **df** | | **p** | |
| Χ² |  | 5.195 |  | 1 |  | 0.023 |  |
| N |  | 50 |  |  |  |  |  |
|  | | | | | | | |

You can see that the statistical significance value (i.e., *p*-value) is **.023** (i.e., *p* = .023). If *p* < .05, you have a statistically significant result, whereas if *p* > .05, you do not have a statistically significant result. Our *p*-value of .023 is less than .05 (i.e., *p* = .023 satisfies *p* < .05). Therefore, we have a statistically significant result; that is, there is a statistically significant association between our two dichotomous variables.

You could report this as follows: A chi-square test for association was conducted between gender and preference for performing competitive sport. All expected cell frequencies were greater than five. There was a statistically significant association between gender and preference for performing competitive sport, χ2(1) = 5.195, *p* = .023.

You can also report any numbers or percentages from the crosstabulation table that you feel are appropriate to explain your results.

## **Reporting All Together:**

A chi-square test for association was conducted between gender and preference for performing competitive sport. All expected cell frequencies were greater than five. There was a statistically significant association between gender and preference for performing competitive sport, χ2(1) = 5.20, *p* = .023. There was a moderately strong association between gender and preference for performing competitive sport, ɸ = .322.