Statistics 101

Proudly built with R, Quarto and GitHub Copilot

Daniel Sánchez Pazmiño

Table of contents

SPSS II: Further statistical inference methods using SPSS

Dependent samples t-test

As you might remember, a dependent samples t-test is used to compare the means of two related groups. In this case, the two groups are related because they are the same participants measured at two different times. The dependent samples t-test is also known as a paired samples t-test or a repeated measures t-test.

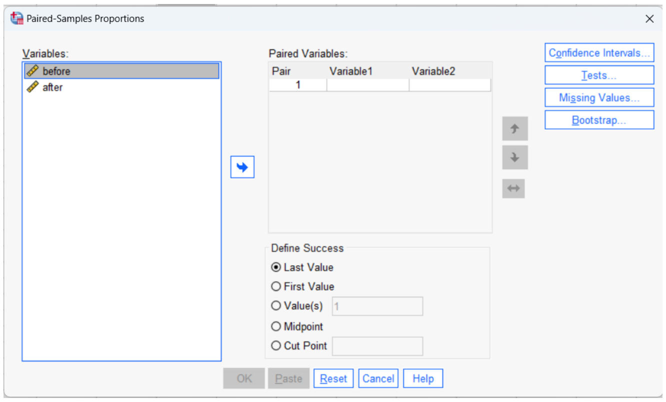
The data for this type of test must follow the following structure:

Participant	Measure before	Measure after
1	10	12
2	11	13
3	12	14
4	13	15

You can see that the data is structured in such a way that each participant has two scores, one for the measure before and one for the measure after. The dependent samples t-test is used to determine whether the mean difference between the two measures is zero or not. In other words, it tests whether the mean of the differences between the two measures is significantly different from zero.

It is not necessary to have the participants identified with a column, but it is important to have the two measures in separate columns. Also remember that paired samples require the same amount of participants in the two groups.

To run the test, go to `Analyze > Compare Means > Paired-Samples T Test`. In the dialog box, select the two variables that you want to compare. In this case, we will select `Measure before` and `Measure after`.



Define Success

● Last Value
○ First Value
○ Value(s) 1
○ Midpoint
○ Cut Point

Figure 1: SPSS Welcome Screen

We will look at the p-value in the `Sig. (2-tailed)` column. If the p-value is less than .05, we can conclude that the mean difference between the two measures is significantly different from zero. In this case, the p-value is .000, which is less than .05, so we can conclude that the mean difference between the two measures is significantly different from zero.

Z-tests for the mean

Z-tests for one and two samples, as you may remember, are used to test hypotheses about population means when the population standard deviation is known. The z-test for one sample is used to test hypotheses about a population mean when the population standard deviation is known. The z-test for two samples is used to test hypotheses about the difference between two population means when the population standard deviation is known.

Unfortunately, SPSS does not offer a direct way to run z-tests. This is because we almost never use a z-test for hypothesis testing about means. While obviously the z-test is present for proportions, for means only the t-test for one and two samples is used. This is because only rarely we have information about the population standard deviation, and any reputable research journal will not believe that you really have information about the population std. dev.

However, if you really want to run a z-test, you can do it partially by hand. We calculate the z-score and then use spss to calculate the critical value or the p-value. Remember that the critical value is calculated based on the significance value that you set, while the p-value is calculated based on the z-score that you obtained. The formula for the z-score is:

$$z = \frac{\bar{x} - \mu}{\sigma}$$

Where $\bar{x}$ is the sample mean, μ is the population mean, and σ is the population standard deviation.

So, if we had a simple, one sample z-test like the following:

> A researcher wants to test whether the mean height of a population is 1.75 meters. She takes a sample of 100 people and finds that the mean height of the sample is 1.73 meters. The population standard deviation is 0.1 meters. Is the mean height of the population 1.75 meters?

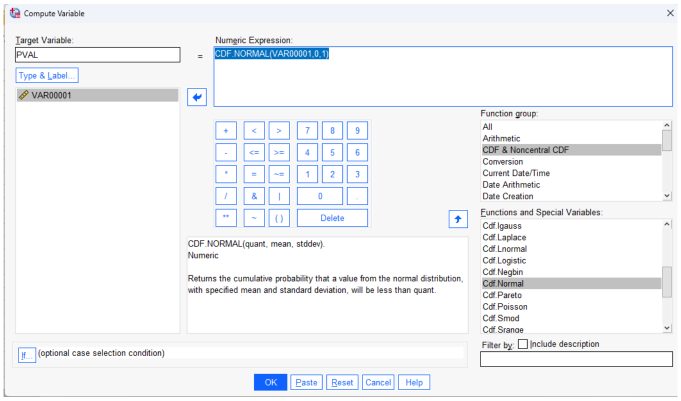
The hypotheses will be written as:

$$H_0 : \mu = 1.75$$
$$H_1 : \mu \neq 1.75$$

We can calculate the z-score by hand:

$$z = \frac{\bar{x} - \mu}{\sigma} = \frac{1.73 - 1.75}{0.1} = -0.2$$

Now, we can use SPSS to calculate the critical value or the p-value. To do this, we will need to input the -0.2 z-score in the Data Editor, and then use the Transform menu to get the cumulative probability.



Compute Variable

Target Variable: PVAL

Numeric Expression: CDF.NORMAL(VAR00001,0,1)

Function group: CDF & Noncentral CDF

Functions and Special Variables: Cdf.Normal

CDF.NORMAL(quant, mean, stddev).
Numeric

Returns the cumulative probability that a value from the normal distribution, with specified mean and standard deviation, will be less than quant.

Figure 2: SPSS Probability Calculator

The result of this was 0.42. This is *left* tailed probability, as mentioned by the SPSS formula guide in the figure. We need to input the variable where we inputted the z-score, as well as the mean and the standard deviation. Since we've already calculated a standardized value, which is the z-score, we need to write down 0 and 1 as the mean and standard deviation.

Because this is a two tailed test, we need to multiply the obtained result of 0.42 by two. So, the p-value is 0.84. Since this is greater than .05, we fail to reject the null hypothesis. We can conclude that the mean height of the population is 1.75 meters.

Chi-square test for goodness of fit

The chi-square tests are one type of several non-parametric statistical inference. If we are running a chi-square test, it means that we are not assuming that the data is normally distributed. The chi-square test for goodness of fit is used to test whether the observed frequencies of a categorical variable differ from the expected frequencies.

The data for this type of test can follow multiple structures, but the most common one is the *long form*. This means that every row represents a participant, and there is a column for the categorical variable that we are interested in. We don't need to calculate frequencies for each category, as SPSS will do this for us.

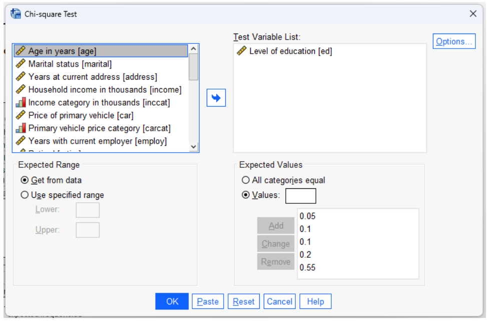
Let us use the `demo` dataset again and let us look at the education. We want to test if the observed frequencies of education differ from the expected frequencies. The expected frequencies are the frequencies that we would expect, and these come from our research.

We might believe that the expected frequencies are 5% for those who did not finish high school, 10% for those who finished high school, 20% for those who finished college, 10% for those with some college, and 55% for those with a graduate degree. Let us use SPSS to test this hypothesis. Write down the hypotheses:

$$H_0 : \text{The observed frequencies are equal to the expected frequencies}$$

$$H_1 : \text{The observed frequencies are not equal to the expected frequencies}$$

To run the test, go to `Analyze > Nonparametric Tests > Legacy Dialogs > Chi-Square`. In the dialog box, select the variable that you want to test. In this case, we will select `Education`. Then, click on `Cells` and select `Expected` and `Observed`.



Chi-square Test

- Age in years [age]
- Marital status [marital]
- Years at current address [address]
- Household income in thousands [income]
- Income category in thousands [inccat]
- Price of primary vehicle [car]
- Primary vehicle price category [carcat]
- Years with current employer [employ]

Test Variable List:

- Level of education [ed]

Expected Range

● Get from data

○ Use specified range

Lower:

Upper:

Expected Values

○ All categories equal

● Values:

0.05
0.1
0.1
0.2
0.55

Figure 3: SPSS Chi-Square Dialog Box

We need to input the variable we're using, which is `ed`, and the expected frequencies. We need to enter them in proportion form, so if the expected frequency is 5%, we need to enter 0.05. As in other cases, we need to enter everything in order based on the information that we have about the labels of the variable in the Variable view.

The results show that the p-value is very small, so we can reject the null hypothesis. We can conclude that the observed frequencies are not equal to the expected frequencies.

Chi-square test for independence

The chi-square test for independence is used to test whether two categorical variables are independent or not. This means that we are testing whether the observed frequencies of one variable are independent of the observed frequencies of the other variable.

We also use the same long form dataset `demo` which has a column for income and education. Let us test the independence of education and income.

The hypotheses are:

$$H_0 : \text{Education and income are independent}$$

$$H_1 : \text{Education and income are not independent}$$

To run the test, go to `Analyze > Descriptive Statistics > Crosstabs`. In the dialog box, select the two variables that you want to test. In this case, we will select `Education` and `Income`. Then, click on `Statistics` and select `Chi-square`.

The results will show several different chi-square tests. The one that we are interested in is the `Pearson Chi-Square` test. The p-value is very small, so we can reject the null hypothesis. We can conclude that education and income are not independent.

ANOVA (Analysis of Variance)

ANOVA is a statistical test that is used to test whether the means of three or more groups are equal or not. The ANOVA test is an extension of the t-test, which is used to test whether the means of two groups are equal or not.

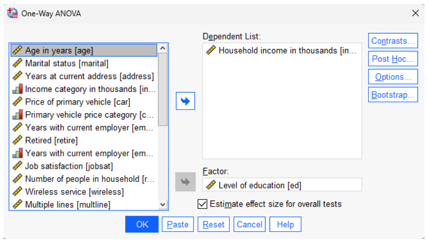
The data for this type of test can once again folow multiple structures, but the most common one is, again, the *long form*. This means that every row represents a participant, and there is a column for the categorical variable that we are interested in as well as quantitative or numerical variable. We don't need to calculate averages for each category, as SPSS will do this for us.

Let us use the `demo` dataset again and let us look at the income. We want to test if the means of income differ between the different levels of education. The hypotheses are:

$$H_0 : \text{The means of income are equal between the different levels of education}$$

$$H_1 : \text{The means of income are not equal between the different levels of education}$$

To run the test, go to `Analyze > Compare Means > One-Way ANOVA`. In the dialog box, select the variable that you want to test. In this case, we will select `Income`.


One-Way ANOVA
Age in years [age]
Marital status [marital]
Years at current address [address]
Income category in thousands [in...
Price of primary vehicle [car]
Primary vehicle price category [c...
Years with current employer [em...
Retired [retire]
Years with current employer [em...
Job satisfaction [jobsat]
Number of people in household [r...
Wireless service [wireless]
Multiple lines [multline]
Contrasts...
Post Hoc...
Options...
Bootstrap...
OK
Paste
Reset
Cancel
Help

Dependent List: Household income in thousands [in...

Factor: Level of education [ed]

☑ Estimate effect size for overall tests

Figure 4: SPSS ANOVA Dialog Box

We need to input the dependent variable we're using, which is `income`, and the factor variable, which is `ed`. The results show that the p-value is very small, so we can reject the null hypothesis. We can conclude that the means of income are not equal between the different levels of education. However, we cannot say which means are different from each other. To do this, we need to run a post-hoc test or a better method, such as regression. Further, because we do not have experimental data, we cannot say that the differences are caused by the different levels of education.

Regression

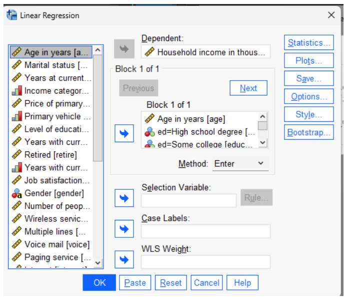
Regression allows to explain the relationship between a dependent variable and one or more independent variables. The dependent variable is the variable that we want to explain, while the independent variables are the variables that we use to explain the dependent variable.

We can run a regression with the exact same data that we used for the ANOVA. We want to explain income, and first let us use age.

$$\text{Income} = \beta_0 + \beta_1 \text{Age} + \epsilon$$

Where β_0 is the intercept, β_1 is the slope, and ϵ is the error term. We are interested in the slope, which is the coefficient of age. SPSS will estimate the intercept and the slope for us, and thus the estimated model will follow the same notation but with a b instead of a β.

To run the test, go to `Analyze > Regression > Linear`. In the dialog box, select the dependent variable that you want to test. In this case, we will select `Income`. Then, click on `Independent(s)` and select `Education`.



Linear Regression

Dependent: Household income in thous...

Block 1 of 1

Age in years [age]
ed=High school degree [...
ed=Some college [educ...

Method: Enter

Selection Variable:

Case Labels:

WLS Weight:

Figure 5: SPSS Regression Dialog Box

The results will show the coefficient on age, which is the most important thing we're interested in. The coefficient on age is 2.147, which we interpret as follows: for every one year increase in age, income increases by 2.147 thousand dollars a year. There is a p-value (Sig.) attached to the coefficient, which is the p-value for the null hypothesis that the coefficient is equal to zero. So, because the p-value is very small, we can reject the null hypothesis. We can conclude that age is a significant predictor of income. However, we cannot say that age causes income to increase, because we don't have experimental data.

How could we make this a multiple regression? We only need to add more independent variables. There will be more coefficients, one for each independent variable, and a p-value.

There is ANOVA analysis of variance table at the bottom of the results. The test that is being done in this ANOVA is a special kind of ANOVA which tells you whether or not all of the variables that you are using to explain the dependent variable are significant. In this case, the p-value is very small, so we can reject the null hypothesis. We can conclude that the variables that we are using to explain income are significant. In the case of a multiple regression, this means that at least one of the independent variables is significant.

Dummy coefficients

We want to explain income using education and age. Because education is not a quantitative variable, we need to tread carefully. If you put the variable `ed` in the list of independent variables in the regression, SPSS will treat it as a quantitative variable. This is not what we want, because education is a categorical variable.

We need to make a prior transformation to the variable `ed`. We need to create dummy variables for each level of education. We will create a dummy variable for each level of education, and we will use the variable `ed` to create the dummy variables. The Transform menu has a function called `Create Dummy Variables`. We need to click on that function and select the variable `ed`.

There will be a dummy variable at the end of the Data Editor Data view. We can use those dummy variables to explain income. However, at most you can put four of those dummy variables in the regression. If you put all five, you will get an error. This is because the fifth dummy variable is redundant, and you need a reference group.

Before trying to transform into different dummy variables, we need to make sure that the variable `ed` is a "Nominal" variable. If it is not, we need to change it to "Nominal" in the Variable View.

When we see the results of this new analysis, and we leave out the did not finish high school dummy variable, we can see that the coefficient on the high school dummy variable is 13.039. This means that the mean income of those who finished high school is 13.039 thousand dollars higher than the mean income of those who did not finish high school. As you can see, this is equivalent to the difference in means that we obtained in the ANOVA, but now we can say which category has a higher mean income.