Statistical tests for SPSS

Paolo Coletti – A.Y. 2010/11 – Free University of Bolzano Bozen

Premise

This book is a very quick, rough and fast description of statistical tests and their usage. It is explicitly designed for an SPSS course and therefore the description of tests is oversimplified and focused only on their practical usage. This book is not indicated for any statistics' course.

Table of Contents

1. Statistical tests

Statistical tests are inference tools which are able to tell us the probability with which results obtained on the sample can be extended to the population. Every statistical test has these features:

- a sample of data $x_1, x_2, \dots, x_n$ and a population, to which we want to extend information and relations found on the sample;
- the null hypothesis H_0 and its contradictory hypothesis H_1;
- prerequisites, special assumptions which are necessary to perform the test;
- the statistic, a function calculated on the data, whose value determines the result of the test;
- significance, also called p-value, from which we can deduct whether accepting or rejecting null hypothesis.

1.1 Example

In order to show all the elements of a statistical test, we run through a very simple example. We want to study the age of workers of a large company. We want to check whether the expected value is 35 years or not.

Of this random variable the only thing we know are the observations on a sample of 100 workers, which are: 25; 26; 27; 28; 29; 30; 31; 30; 33; 34; 35; 36; 37; 38; 30; 30; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 51; 52; 20; 54; 55; 56; 57; 20; 20; 20; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 35; 36; 37; 35; 36; 37; 35; 36; 37; 35; 36; 37; 35; 36; 37; 35; 36; 37; 35; 36; 37; 35.

We now formulate the test's hypotheses:

- H_0: expected value of age $= 35$
- H_1: expected value of age $\neq 35$

Now we calculate the age average on the sample, $\overline{\text{age}}_{100} = 36.2$, which is an estimation for the expected value. We compare this result with the 35 of the H_0 hypothesis and we find a difference of $+1.2$. At this point, we ask ourselves whether this difference is large enough, implying that the expected value is not 35 and thus H_0 must be rejected, or is small and can be caused by an unlucky choice of the sample and therefore H_0 must be accepted.

However, this conclusion in a statistical research cannot be drawn from a subjective decision whether the difference is large or small. It is taken using formal arguments and therefore we must rely on this statistic function

$$\frac{\overline{\text{age}}_n - \text{hypothesized expected value}}{\sqrt{\text{sample variance}/n}},$$

which, on our data, has a value of $+1.40$. The only reason why we have built this statistic instead of using directly the difference at the numerator is because we know the statistic's distribution and we can calculate the significance. Therefore we can calculate that the probability of getting a value larger than $+1.40$ or smaller than -1.40 is 16%. This value is called significance or p-value.

If significance is large it means that, supposing H_0 to be true and taking another random sample, the probability of obtaining a worse result is large and therefore the result that we have obtained can be considered to be really close to 0, something which pushes us to accept the idea that H_0 is true. When, instead, significance is small, it means that if we suppose that H_0 is true we have a small probability of getting such a bad result, something which pushes us to believe that H_0 be false. In the example's situation we have a significance of 16%, which usually is considered large (the chosen cut point is typically 5%) and therefore we accept H_0.

Confidence is equal to 100% minus the significance.

1.2 Accept and reject

At the end of the statistical test we must decide whether accepting or rejecting:

- if significance is above the significance level (usually 5% or 1%), we accept H_0;
- if significance is below the significance level, we reject H_0.

It is very important to underline the fact that, for technical reasons that are not be explained here, when we reject we are almost sure that H_0 is false, since we are keeping errors under a small significance level. However, when we accept we may not say that H_0 be true, since we do not have any estimation on errors. Therefore, rejecting is a sure thing, while *accepting is a "no answer" and from it we are not allowed to draw any conclusion*.

1.3 Parametric and non-parametric test

There are parametric and non-parametric statistical tests. A parametric test implies that the distribution in question is known up to a parameter or several parameters. For example, it is believed that many natural phenomena are normally distributed. Estimating μ and σ^2 of the phenomenon is a parametric statistical problem, because the shape of the distribution, a normal one, is known up to these two parameters. On the other hand, non-parametric test do not rely on any underlying assumptions about the probability distribution of the sampled population. For example, we may deal with continuous distribution without specifying its shape.

Non-parametric tests are also appropriate when the data are non-numerical in nature but can be ranked, thus becoming rank tests. For example, taste-testing foods we can say we like product A better than product B, and B better than C, but we cannot obtain exact quantitative values for the respective measurements. Other examples are tests where the statistic is not calculated on sample's values but on the relative positions of the values in their set.

1.4 Prerequisites

Each test, especially parametric ones, may have prerequisites which are necessary for the statistic to be distributed in a known way (and thus for us to calculate its significance). A typical prerequisite for many parametric tests is that the sample comes from a certain distribution.

2. Tests

2.1 Student's t test for one variable

Prerequisites: none.

H_0: expected value = $\boldsymbol{m}$

SPSS: Analyze → Compare Means → One-Sample T Test

Student's t test is the one we have already seen in the example. It is a test which involves a single random variable and checks whether its expected value is m or not.

2.2 Student's t test for two populations

Prerequisites: two populations A and B and the variable must be distributed normally on the two populations

H_0: expected value on population A = expected value on population B

SPSS: Analyze → Compare Means → Independent-Samples T Test

This test is used whenever we have two populations and one variable calculated on this population and we want to check whether the expected value of the variable changes on the populations.

For example, we want to test

- H_0: $\mathrm{E(height)\ for\ male} \le \mathrm{E(height)\ for\ female}$
- H_1: $\mathrm{E(height)\ for\ male} > \mathrm{E(height)\ for\ female}$

2.3 Student's t test for paired data

Prerequisites: two variables ζ and ξ on the same population and $\zeta - \xi$ must be normally distributed

H_0: $\textbf{expected value of } \boldsymbol{\zeta} - \textbf{expected value of } \boldsymbol{\xi} = \boldsymbol{m}$

SPSS: Analyze → Compare Means → Paired-Samples T Test

This test is used whenever we have a single population and two variables calculated on this population and we want to check whether the expected value of these two variables has a certain difference m or not.

For example, we want to test whether population's income in a country has decreased by 2. We take a sample of 10 people's income and then we take the same 10 subjects' income the next year

Income 2010 (thousands €)	Income 2011 (thousands €)	Difference 2012 – 2010
20	21	+1
23	23	0
34	36	+2
53	50	-3

43	40	-3
45	44	-1
36	12	-24
76	80	+4
44	45	+1
12	15	+3

The subjects must be exactly the same.

Hypotheses are:

- H_0: $\mathrm{E(income)for\ 2011 - E(income)for\ 2010} = -2$
- H_1: $\mathrm{E(income)for\ 2011 - E(income)for\ 2010} \neq -2$

2.4 One-way analysis of variance (ANOVA)

Prerequisites: k populations, variable is normally distributed on every population with the same variance

H_0: expected value of the variable is the same on all populations

SPSS: Analyze → Compare Means → One-Way ANOVA

This test is the equivalent of Student's t test for two unpaired populations when the populations are more than two. We note that if only one population has an expected value different from the other, the test rejects. Therefore, a rejection guarantees us that populations do not have the same expected value but does not tell us which populations are different and how.

For example, we have heights for young (180; 170; 150; 160; 170), adults (170; 160; 165) and old (155; 160; 160; 165; 175; 165) and we want to check

- H_0: $\mathrm{E(height)for\ young} = \mathrm{E(height)\ for\ adults} = \mathrm{E(height)\ for\ old}$
- H_1: at least one of the $\mathrm{E(height)}$ is different from the others

2.5 Kolmogorov-Smirnov test

Prerequisites: none.

H_0: variable follows a known distribution

SPSS: Analyze → Nonparametric Tests → 1-Sample K-S

This is a rank test which checks whether a variable is distributed, on the population, according to a known distribution specified by the researcher.

2.6 Sign test

Prerequisites: continuous distribution.

H_0: median is m

SPSS: Analyze → Nonparametric Tests → Binomial

Sign test is a rank test which tests the central tendency of a probability distribution. It is used to decide on whether the population median equals or not the hypothesized value.

Consider the example when 8 independent observations of a random variable X having a continuous distribution are 0.78, 0.51, 3.79, 0.23, 0.77, 0.98, 0.96, 0.89. We have to decide whether the distribution median is equal to 1.00. We formulate the two hypotheses:

- H_0: $\text{median} = 1.00$
- H_1: $\text{median} \neq 1.00$

2.7 Mann-Whitney (Wilcoxon rank sum) test

Prerequisites: the two probability distributions are continuous

H_0: position of distribution for population A = position of distribution for population B

SPSS: Analyze → Nonparametric Tests → 2 Independent Samples

Suppose two independent random samples are to be used to compare two populations and we are unwilling to make assumptions about the form of the underlying population probability distributions (and therefore we cannot perform Student's t test for two populations) or we may be unable to obtain exact values of the sample measurements. If the data can be ranked in order of magnitude, the Mann-Whitney test (also called Wilcoxon rank sum test) can be used to test the hypothesis that the distributions of the two populations have the same position.

2.8 Wilcoxon signed rank test

Prerequisites: the difference is a random variable having a continuous probability distribution.

H_0: position of distribution for variable 1 = position of distribution for variable 2

SPSS: Analyze → Nonparametric Tests → 2 Related Samples

Rank tests can also be employed to compare two probability distributions when a paired difference design is used. For example, consumer preferences for two competing products are often compared by analyzing the responses in a random sample of consumers who are asked to rate both products. Thus, the ratings have been paired on each consumer.

2.9 Kruskal-Wallis test

Prerequisites: there are 5 or more measurements in each sample; the k probability distributions from which the samples are drawn are continuous

H_0: distribution of populations is the same

SPSS: Analyze → Nonparametric Tests → K Independent Samples

The Kruskal-Wallis test is the Mann-Whitney test when more than two populations are involved. Its corresponding parametric test is the Analysis of Variance.

For example, a health administrator wants to compare the unoccupied bed space for three hospitals. She randomly selects 10 different days from the records of each hospital and lists the number of unoccupied beds for each day:

Hospital 1	Hospital 2	Hospital 3
6	34	13
38	28	35
3	42	19
17	13	4
11	40	29

30	31	0
15	9	7
16	32	33
25	39	18
5	27	24

We test

- H_0: the distributions of the number of unoccupied beds are in the same positions for all three hospitals
- H_1: at least one of the hospitals has distribution in a different position with respect to the others.

2.10 Pearson's correlation coefficient

Prerequisites: coupled data

H_0: $\mathrm{Corr}(X,Y)=0$, i.e. X and Y are linearly independent

SPSS: Analyze → Correlate → Bivariate

Consider two random variables, X and Y, of which we have only n couples of outcomes, $(x_i; y_i)$. It is important that the outcomes that we have are in couples, since we are interesting in estimating the correlation between the two variables. We use as estimator the Pearson's correlation coefficient which is defined as

$$r == \frac{\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\cdot\sum_{i=1}^{n}(y_i-\bar{y}_n)^2}}.$$

The value of r must lie between -1 and $+1$, independently from how large or small are the numbers x_i and y_i. A value of r near or equal to zero is interpreted as little or no correlation between X and Y. In contrast, the closer r comes to -1 or $+1$, the stronger is the correlation of these variables. Positive values of r imply a positive correlation between X and Y. That is, if one increases, the other one increases as well. Negative values of r imply a negative correlation. In fact, X and Y move in the opposite directions: when X increases, Y decreases and vice versa. In sum, this coefficient of correlation reveals whether there is a common tendency in moves of X and Y.

For example, suppose we have these 11 couples of data

$\boldsymbol{x}$	2	3	4	3	5	6	7	3	1	3	4
$\boldsymbol{y}$	5	5	7	5	7	7	14	5	3	1	12

we get $r=0.732$ with 11 couples of data and the value of our statistic is 3.223 and a significance of 1.04%. Therefore, taking a significance level of 5%, 11 couples of data with Pearson's correlation coefficient of 0.732 are enough to prove that the correlation is different from 0 and therefore the two variables are not independent.

2.11 Spearman's rank correlation coefficient

Prerequisites: coupled ranked data or coupled data from continuous distributions

H_0: ranks are the same

SPSS: Analyze → Correlate → Bivariate

The Spearman's rank correlation coefficient is the non parametric version of the Pearson’s correlation coefficient.

2.12 Chi-square test for a one-dimensional contingency table

Prerequisites: at least 5 elements per cell

H$_0$: classification follows a predetermined distribution

SPSS: Analyze → Nonparametric tests → Chi-Square

For example, suppose a large supermarket chain conducts a consumer preference survey by recording the brand of bread purchased by customers in its stores. Assume the chain carries three brands of bread, A, B and C. The brand preferences of a random sample of 150 consumers are observed, and the resulting count data are as follows: A: 61, B: 53, C: 36. Do these data indicate that a preference exists for any of these brands?

We build a table of observed counts and a table of predicted counts

A	B	C
61	53	36

observed counts

A	B	C
50	50	50

predicted counts
under H$_0$ hypothesis

The test statistic is the table’s chi square, a measure calculated as $\chi^2 = 6.52$ with a significance of 3.84% and therefore we reject, meaning that consumers’ preferences are not uniform and that there is at least one type of bread that has a probability different from $1/3$.

As another example, using the same data we want to check whether the bread’s probabilities follow a 40%, 40%, 20% distribution:

- H$_0$: $p_1 = 40\%$ and $p_2 = 40\%$ and $p_3 = 20\%$
- H$_1$: $p_1 \neq 40\%$ or $p_2 \neq 40\%$ or $p_3 \neq 20\%$.

The observed and predicted tables are

A	B	C
61	53	36

observed counts

A	B	C
60	60	30

predicted counts
under H$_0$ hypothesis

and $\chi^2 \approx 1.43$ with a significance of 48.8%. Therefore we accept, meaning that our sample is not able to prove that consumers’ preference is not 40%, 40%, 20%.

2.13 Chi-square test for a two-dimensional contingency table

Prerequisites: at least 5 elements per cell

H$_0$: classifications are independent

SPSS: Analyze → Descriptive Statistics → Crosstabs → Statistics → Chi-square

Suppose, for example, that an automobile magazine is interested in determining the relationship between the size and manufacturer of newly purchased automobiles. One thousand recent buyers of cars made in Germany are randomly sampled, and each purchase is classified with respect to the size (small, intermediate, and large) and manufacturer of the automobile (Volkswagen, BMW, Opel, Mercedes). The data are summarized in the two-way table:

Size\Manufacturer	**VW**	**BMW**	**Opel**	**Mercedes**	**Totals**
Small	157	65	181	10	413
Intermediate	126	82	142	46	396
Large	58	45	60	28	191
Totals	341	192	383	84	1000

This table is called a contingency table. It presents multinomial count data classified in two dimensions, namely automobile size and manufacturer. We test whether the two classifications, manufacturer and size in our example, are independent.

- H_0: row variable and column variable are independent
- H_1: row variable and column variable are independent

We use the chi square statistic to compare the observed and predicted counts in each cell of the contingency table $\chi^2 \approx 45.81$, significance is 0.000003%: therefore we reject the hypothesis of independence and we conclude that the size and manufacturer of a car selected by a purchaser are dependent.