

# **CHANCE ENCOUNTERS**

## **Making Sense of Hypothesis Tests**

Howard Fincher

Learning Development Tutor

Upgrade Study Advice Service

Oxford Brookes University

## PREFACE

This guide has a restricted aim.

This guide will not tell you about the design of experiments or surveys, or about how to use Excel/SPSS. Nor is it a guide about which hypothesis test to choose. And there are no worked examples.

**This guide is for students following a course with an introductory statistics component, who find themselves directed towards performing a hypothesis test with Excel/SPSS.**

Such students are often trying to acquire knowledge in three areas simultaneously. First, there is knowledge about the theory of all hypothesis tests. Second, there is the detailed knowledge about one (or more) particular hypothesis test(s). Third, there is the specific knowledge about how to perform such a test with Excel/SPSS software.

This guide only covers the first of these branches of knowledge. It aims to give you an overview of the framework of thought that surrounds all hypothesis tests, and seeks to introduce you to the philosophy of chance that lies behind all hypothesis tests.

A dictionary will suggest that one of the meanings of ‘chance’ is ‘the probability of something happening’. And the philosophy of chance that lies behind all hypothesis tests is that which has been codified in the mathematics of probability.

This guide explains the background that might assist you to write some meaningful and sensible comments about the “p-value” that Excel/SPSS gives you at the end of a z test, or t test, or F test, or  $\chi^2$  test or (almost) any other hypothesis test.

## ACKNOWLEDGEMENTS

I have been inspired by many teachers in the past. I hope that the best of their work has found a place in this guide. The responsibility for shortcomings remains with me.

Any comments on this text, or suggestions for improvement of the text, can be made to the author at [upgrade@brookes.ac.uk](mailto:upgrade@brookes.ac.uk)

## INTRODUCTION

A hypothesis test is not a calculation. A hypothesis test involves many calculations, but the hypothesis test itself is not a calculation.

When an Excel/SPSS hypothesis test concludes with a p-value of, say,  $p = 0.0123$  or  $p = 0.0789$ , it gives the misleading impression that a hypothesis test is a calculation.

This guide invites you to think of a hypothesis test as a process. This guide would like you to think of a hypothesis test as a carefully nuanced, thoughtfully designed and logically ordered decision-making process whose myriad associated numerical calculations are hidden from view in the Excel/SPSS software.

This guide describes what is going on in the background as the Excel/SPSS software performs a hypothesis test. This guide describes eight components that have to come together in a logical order to create the decision making process called a hypothesis test.

These eight components are:

**The Foundation**

**The Null Hypothesis**

**The Alternative Hypothesis**

**The Significance Level of the Test**

**The Key Statistic for the Test**

**The p-value of the Data**

**The Decision**

**The Possibility of an Erroneous Decision**

This step-by-step guide tells you about the place and the function of these eight components in the process, and how each component provides you with a contribution that will help you to write meaningful and sensible comments about the p-value in the Excel/SPSS output.

## THE FOUNDATION

A hypothesis test is often called a significance test. Let's explore why this is.

A researcher who is recording some interesting experimental data will want to be able to distinguish whether the interesting results are due to random chance variation or are due to a special factor (or significant effect) in the research situation.

The foundation of a hypothesis test is the researcher's decision to reduce a real research situation to an imaginary competition between two rival hypotheses.

The reduction of a real research situation to an imaginary competition between two rival hypotheses is not automatic. (A complex research situation can be reduced to two competing hypotheses in several different ways, and each of these reductions can be the beginning of a fresh hypothesis test.)

When such a reduction has been made, a hypothesis test is the name given to a rational way of choosing between these two rival hypotheses.

These two rival hypotheses are referred to as the 'Null Hypothesis' (or  $H_0$  or NH) and the 'Alternative Hypothesis' (or  $H_1$  or AH).

*[This guide will refer to the 'Null Hypothesis' as the NH and the 'Alternative Hypothesis' as the AH.]*

As you read through this guide, you will notice that a hypothesis test does not give equal weight to each hypothesis. The two rival hypotheses are not on an equal footing. Be aware that there is an asymmetry here.

In fact, a hypothesis test places most of its weight on the NH. And that is because it is only the NH that is tested. So the decision at the conclusion of a hypothesis test is always expressed in relation to the NH. This decision will be of the form 'do not reject the NH' or of the form 'reject the NH in favour of the AH'.

Since a real research situation is reduced to an imaginary competition between two rival hypotheses, it is important to realise that either of the two rival hypotheses may be true, or neither of them may be true.

### CONTRIBUTION 1

Either the clause "**do not reject the null hypothesis**" or the clause "**reject the null hypothesis in favour of the alternative hypothesis**" will contribute to your paragraph of interpretation of the p-value at the conclusion of an Excel/SPSS hypothesis test.

## THE NULL HYPOTHESIS

A hypothesis test gives special (almost exclusive) attention to the NH because this is the hypothesis that is being tested.

So which of the two imaginary rival hypotheses do you choose as the NH?

Of the two rival hypotheses, you choose as the NH the hypothesis that expresses the idea of 'no difference', 'no change', 'no effect', 'no correlation', or 'no association'. The NH expresses the idea that there is nothing interesting happening in your research situation, and there are no special factors at work.

When you choose your NH like this, your NH is usually the opposite of what you would like to believe is the case.

Your NH has to be chosen like this so that your reduced research situation can be analysed from the point of view of random chance variation.

Your NH is chosen in this way so that your experimental research data has the opportunity to contradict and to discredit it. Remember that it is only the NH that is being tested.

### CONTRIBUTION 2

Because a hypothesis test only tests the NH, **a sentence concisely describing your NH** will contribute to your paragraph of interpretation of the p-value at the conclusion of an Excel/SPSS hypothesis test.

## THE ALTERNATIVE HYPOTHESIS

The AH is not tested. The AH is outside the central process of the hypothesis test.

Having been given a role at the very beginning of the hypothesis test process, the AH waits off stage, waits in the wings, and will only appear again at the end of the hypothesis test if the NH is rejected.

However, the AH does exercise an unseen influence over the hypothesis test process.

Because of the way that the NH has been chosen, the AH expresses the idea that there is 'some difference', 'some change', 'some effect', 'some correlation', or 'some association'. The AH expresses the idea that there is something interesting happening in your research situation, and there is a special factor at work.

When you choose your AH like this, your AH is usually what you would like to believe is the case.

Your AH is your research hypothesis, so take your time over your choice of an AH.

If you choose an AH that implies any change away from the NH, then your choice makes the hypothesis test a 'two tailed' test (sometimes known as a 'two sided' test).

If you choose an AH that implies a definite increase from the NH, then your choice makes the hypothesis test a 'right tailed' test.

Similarly, if you choose an AH that implies a definite decrease from the NH, then your choice makes the hypothesis test a 'left tailed' test.

A 'right tailed' test and a 'left tailed' test are collectively known as 'one tailed' tests (and sometimes as 'one sided' tests).

It is not sufficient to rely on the experimental data results when you are choosing an AH that implies a 'one tailed' test. You need to be able to justify choosing a 'one tailed' test on subject-specific grounds rather than just numerical grounds.

Unless directed otherwise, this guide suggests that you to choose an AH that leads to a 'two tailed' test. The exception to this advice is that if you are performing an (ANOVA) F test or a  $\chi^2$  test you will normally perform a 'right tailed' test for a 'two sided' AH.

### CONTRIBUTION 3

Because a hypothesis test only tests the NH, **a sentence concisely describing your AH** will contribute to your paragraph of interpretation of the "the p-value" at the conclusion of an Excel/SPSS hypothesis test as evidence for, and as an indication of, whether the test is **a two tailed test**, or **a right tailed test**, or **a left tailed test**.

## THE SIGNIFICANCE LEVEL OF THE TEST

You have to choose a significance level for your test. This significance level is also known as the alpha-value.

The most common significance level in use is 5%. (You may also be directed to use a significance level of 1% or 0.1%.)

The significance level of 5% is equivalent to the alpha-value,  $\alpha = 0.05$ .

*[This guide assumes that you are familiar with the equivalence of percentages and decimal fractions.]*

*Remember, any percentage can be converted to a decimal (fraction) by division by 100, and any decimal (fraction) can be converted to a percentage by multiplication by 100.]*

On the basis of your NH (which is an imaginary model of the real research situation), the Excel/SPSS software imagines all possible random samples that could be recorded and puts them into two classes. There is the class of samples with the smaller overall differences due to random chance variation, and there is the class of samples with the larger overall differences due to random chance variation.

When you choose a significance level for your test, you are choosing the position of the boundary between these two classes. A significance level of 5% means that you are going to put the boundary between a (large) class containing 95% of imaginary samples with the smaller overall differences and a (small) class containing 5% of imaginary samples with the larger overall differences.

A little later in the process, you will see that the hypothesis test will want to know to which of these two classes your experimental data belongs.

Your choice of a significance level for the test is your way of defining the criterion with which you are going to make a decision about the NH. (Remember that your NH is an imaginary hypothesis which may or may not be true.)

### CONTRIBUTION 4

The phrase “**a significance level of 5% ( $\alpha = 0.05$ )**” will contribute to your paragraph of interpretation of the “the p-value” at the conclusion of an Excel/SPSS hypothesis test.

## THE KEY STATISTIC FOR THE TEST

As hinted in the last section, the Excel/SPSS software imagines all possible random samples that could be recorded based upon your choice of a NH.

This work is at the heart of the hypothesis test.

The hypothesis test uses a key statistic such as  $z$ , or  $t$ , or  $F$ , or  $\chi^2$ , etc, to refer to a complex mathematical chance variation model to generate all possible random samples that could be recorded based upon your choice of a NH. (The details of your research situation will suggest which one of these key statistics will provide you with a suitable model.)

Numerical values of  $z$ , or  $t$ , or  $F$ , or  $\chi^2$ , etc, encode the overall difference between random samples and the NH.

Probabilities are now associated with each numerical value of the key statistic for your test.

The choice of a 5% significance level for the test means that the hypothesis test is tracking (with a numerical value of your key statistic) the position of the boundary between the 95% of samples with the smaller overall differences and the 5% of samples with the larger overall differences in relation to the NH.

### CONTRIBUTION 5

The identification of the key statistic for your test, such as '**a z test**', or '**a t test**', or '**an F test**', or '**a  $\chi^2$  test**' will contribute to your paragraph of interpretation of the p-value at the conclusion of an Excel/SPSS hypothesis test.

## THE p-VALUE OF THE DATA

This is the opportunity for your experimental data to contest the imaginary NH.

Excel/SPSS reads your experimental data and works out a numerical value of the key statistic. This numerical value of  $z$ , or  $t$ , or  $F$ , or  $\chi^2$ , etc, is a numerical measure of the overall difference between your experimental data and the NH.

Excel/SPSS now assigns a probability to your data value of  $z$ , or  $t$ , or  $F$ , or  $\chi^2$ , etc. And the probability it assigns is the probability that a numerically larger value of  $z$ , or  $t$ , or  $F$ , or  $\chi^2$ , etc, than your data value of  $z$ , or  $t$ , or  $F$ , or  $\chi^2$ , etc, could be expected to occur on the basis that the NH is true.

This probability is the p-value.

This p-value is the probability of obtaining a key statistic value for your data as different, or more different, from that envisaged by the NH.

This p-value of your experimental data is also sometimes known as the significance level of the data. (Remember that the alpha-value is the significance level of the test.)

*[Excel presents very small values of the p-value in an 'informal' version of scientific notation such as 2.34E-05. This p-value is  $p = 0.0000234$ .*

*SPSS presents all very small values of the p-value as 0.000]*

### CONTRIBUTION 6

The '**p-value**' must be stated at the conclusion of an Excel/SPSS hypothesis test.

## THE DECISION

Because the hypothesis test is a test of the NH, the p-value is always constructed on the assumption that the NH is true.

The decision you are about to make is a decision about the (imaginary) NH and not about the real experimental data. You have been using the experimental data to test the NH, and not the other way round.

A significance level for the test of 5% means that your method of testing the NH is only to reject the NH if the probability of an overall difference between your experimental data (encoded in a value of z or t or F or  $\chi^2$ , etc) and the NH, or a greater overall difference, occurring by random chance variation is less than 0.05.

By using a 5% significance level, you are deciding to reject the NH for those differences that random chance variation alone (based on the NH) would explain as occurring less than 5% of the time.

To make your decision, you compare the p-value (the significance level of the data) with the alpha-value (the significance level of the test).

- If the p-value is greater than the alpha-value ( $\alpha = 0.05$ ), then your decision is 'do not reject the NH' and your conclusion is that the result of the test is 'not statistically significant'.

A 'do not reject the NH' decision does not imply or mean or prove that the NH is true. Such a decision only suggests that there is not sufficient research experimental evidence against the NH. In other words, there was a difference/correlation/association, but the difference/correlation/association was not large enough to decide against random chance variation as an explanation of this difference/correlation/association.

- If the p-value is less than the alpha-value ( $\alpha = 0.05$ ), then your decision is 'reject the NH in favour of the AH' and your conclusion is that the result of the test is 'statistically significant'.

A 'reject the NH in favour of the AH' decision does not imply or mean or prove that the AH is true. Such a decision only suggests that the AH may be true. In other words, the difference/correlation/association was great enough to decide against random chance variation (based on the NH) as an explanation of this difference/correlation/association.

*[This guide advises you against using such terms as 'accept the NH', or 'reject the AH', or 'accept the AH'.]*

### CONTRIBUTION 7

Your paragraph of interpretation of the p-value at the conclusion of an Excel/SPSS hypothesis test will **compare your p-value for the data with the alpha-value for the test**, move on to **an explicit statement of your decision about the NH** and use the phrase either **"not statistically significant"** or **"statistically significant"**.

## THE POSSIBILITY OF AN ERRONEOUS DECISION

This component may not seem to be part of the hypothesis test. However, an appreciation of the quality of the decision you make can contribute to your concluding remarks, and to your overall understanding of the hypothesis test process.

First, there are several assumptions associated with each particular hypothesis test. And although hypothesis tests have a certain robustness built into them, the validity of the test becomes an issue when there are serious breaches of any of these assumptions.

Second, the decision that has been made is not without the possibility of mistake. In fact, the internal logic of the components of a hypothesis test means that mistakes can sometimes be expected to be made.

To take this idea of a mistaken decision further, we need to visit error management theory. Here the discourse is about 'false positive' and 'false negative' errors.

A 'false positive' error occurs when the researcher rejects the NH when the NH represents the true state of affairs in the real world.

A 'false negative' error occurs when the researcher does not reject the NH when the NH does not represent the true state of affairs in the real world.

How does this work out in practice? Consider a spam filter as a decision-making hypothesis test.

NH: My next e-mail is not spam.

AH: My next e-mail is spam.

A 'false positive' error occurs when the spam filter intercepts your next e-mail when the true state of affairs in the real world is that your next e-mail is not spam.

A 'false negative' error occurs when the spam filter does not intercept your next e-mail when the true state of affairs in the real world is that your next e-mail is spam.

In hypothesis tests, it is traditional to name a 'false positive' error as a Type I Error, and a 'false negative' error as a Type II Error.

Sadly, these two kinds of errors are inversely related (to a greater or lesser extent). As you lower the risk of one of them, the risk of the other rises.

In most contexts, it is generally agreed that a Type I Error is more serious.

### CONTRIBUTION 8

Your paragraph of interpretation of the p-value at the conclusion of an Excel/SPSS hypothesis test should **indicate an awareness of the quality of the decision made.**

## MAKING SENSE OF HYPOTHESIS TESTS

When Excel/SPSS gives you a p-value at the end of a z test, or t test, or F test, or  $\chi^2$  test or any other test, you will normally be expected to write a paragraph of interpretation of the p-value. You should consider writing complete sentences with the minimum of numerals and symbols. And you should remember that the focus of this paragraph is the test, and not the real world which produced the experimental data.

- Here is a model paragraph when the p-value is, say, 0.0123 and you decide to reject the NH in favour of the AH. Make a choice or make an addition at [ ] or ( ).

**With a 5% significance level [z or t or F or  $\chi^2$ ] test ( $\alpha = 0.05$ ) of the null hypothesis [specify it] against the alternative hypothesis [specify it] (two tail test or right tail test or left tail test), the data significance level was computed by [Excel or SPSS] as  $p = 0.0123$ .**

**Since  $p = 0.0123 < \alpha = 0.05$ , I reject the null hypothesis in favour of the alternative hypothesis.**

**I conclude that the experimental data is statistically significantly different from the null hypothesis, and may be consistent with the alternative hypothesis.**

**The mathematics of chance suggests that this hypothesis test will reach the correct conclusion on 95% of the occasions when this test is correctly used in this research situation, and when attention is paid to the assumptions on which the test is based.**

- Here is a model paragraph when the p-value is, say, 0.0789 and you decide not to reject the NH. Make a choice or make an addition at [ ] or ( ).

**With a 5% significance level [z or t or F or  $\chi^2$ ] test ( $\alpha = 0.05$ ) of the null hypothesis [specify it] against the alternative hypothesis [specify it] (two tail test or right tail test or left tail test), the data significance level was computed by [Excel or SPSS] as  $p = 0.0789$ .**

**Since  $p = 0.0789 > \alpha = 0.05$ , I do not reject the null hypothesis.**

**I conclude that the experimental data is not statistically significantly different from the null hypothesis, and may be consistent with it.**

**The mathematics of chance suggests that this hypothesis test will reach the correct conclusion on 95% of the occasions when this test is correctly used in this research situation, and when attention is paid to the assumptions on which the test is based.**

## FREQUENTLY ASKED QUESTIONS

### **Is the p-value at the conclusion of a hypothesis test affected by the size of my sample of data?**

Yes, the p-value is sensitive to the sample size. The smaller your sample the easier it is 'to discover' a statistically significant result. Larger samples have the potential to lead to more credible results.

### **Is a statistically significant result a credible result?**

Not necessarily. The credibility attached to any particular p-value depends on the way that the research was carried out. In general, the larger the sample and the better the design of the study, the more credible will be your result.

### **What is the credibility of the result of my hypothesis test?**

The credibility of your test result is the percentage probability that the largest, most perfectly designed and executed study possible would produce the same outcome.

### **Is a statistically significant result actually significant in the real world?**

Probably not, and for three reasons. First, it is probably the case that your data can always be associated with a statistically significant result by adjusting the significance level of the test. Second, your hypothesis test is based on a reduction of the real world to an imaginary competition between two rival hypotheses. Third, your data is consistent with an infinite set of other null hypotheses that you have not considered, and inconsistent with yet another infinite set of null hypotheses that you not considered.

### **Would I achieve the same outcome ['a statistically significant result' or 'a not statistically significant result'] if I had used a 'Confidence Interval' approach to my data?**

Yes. When both approaches are appropriate and feasible (and the significance level of the test is matched with the confidence level), a 'hypothesis test' approach and a 'confidence interval' approach using your data will come to the same conclusion. The latter approach is often more useful because it shows at a glance the infinite set of null hypotheses which would lead to 'a not statistically significant result' and the infinite set of null hypotheses which would lead to 'a statistically significant result'.