# Understanding Statistical Significance - Read this first

# If you don’t have a basic understanding of how statistics are used to assess a test result you will not be able to use the Excel tool you can download here. (confbt)

I’m going to explain how we use statistics to answer these questions:

* What size of sample should I use for a mailing or email test?
* How will I know whether a test result is reliable?
* How can I keep testing costs down whilst still finding the answers I need?

**NOTE** You do nothave to be a maths genius to be a good direct marketer; but you do need to understand that test evaluation cannot be done without the use of statistics.

I’ll take you through the calculations required to answer the above questions – this will help you understand the process. Once you have worked through a few examples you’ll be able to use the Excel tool with confidence.

## Sample sizes

Testing can help you make sound judgments, but it is not an exact science. Testing helps you identify the option that offers the **greatest probability of success**.

That word probability is very important - a test result does not guarantee that when repeated the same activity will produce the same result. It is therefore necessary for you to understand a little about statistics. This is so that:

1. **you will know how many people you need in a sample for a test result to be ‘significant’, and thus be a reliable predictor of what will happen when you ‘roll-out’ the chosen mailing or email to the full list**
2. **you will be able to decide, after the event, whether a test result can be relied upon i.e. if it is a ‘significant’ result**

**NOTE – measuring the statistical validity of a test with the formulae we are going to introduce shortly requires that we know the size of samples in each test cell. So these formulae are only used when evaluating tests of mailings and emails.**

**When we are evaluating results from online ads, newspapers, TV commercials and so on we use a different method which will be explained later.**

Direct mail and email tests are conducted on samples from the various lists available. For comparative tests to be reliable you must ensure that you are testing 'like with like' - in other words the samples used for each test cell must be:

1. **Matched with each other in terms of composition and characteristics**
2. **The same size, or at least each of a known size so that you can allow for
size variances in evaluation**
3. **Randomised, so that they are entirely typical of the universe they represent, so that you can predict the eventual performance of a 'roll-out' from the test data**
4. **Large enough to give a 'statistically significant' number of responses**

## Randomisation

To ensure that samples are truly representative of their total ‘universe’, they would typically be chosen on a systematic basis i.e. 1 in 'n' samples. (For example, to produce a sample of 10,000 names from a list of 300,000 you would select every 30th name - this ensures randomisation across the list, eliminating bias caused by keeping the list in say, chronological or geographical order).

**NOTE - This is a very important point - the simplest and quickest way of extracting 10,000 names from a list of 300,000 would be to take the first 10,000 records. If you did this with a list held in chronological order, you would pick the 10,000 newest or oldest names, which would not be typical of the list as a whole.**

There are three basic statistical concepts involved in planning tests:

1. **Confidence level** - This is the number of times out of 100 that one could expect the test result to be repeated in a ‘roll-out’ - the levels commonly used in direct marketing testing are between 80% and 95%.
2. **Limit of Error** - Statistics is not an exact science and every test result is subject to a plus or minus correction defined by sample size and response level.
3. **Significance** - this means simply - is the difference observed sufficiently large for it to be outside the variances expected due to the limits of error? If the answer is ‘yes’ the result is significant, if ‘no’ it is not significant.

## Selecting Samples for Testing

The minimum sample size will vary according to how precise you want to be. To repeat, testing relies on the laws of statistics and these are not exact. To ‘read’ a test result with confidence we need at least a basic knowledge of how statistics work.

Our first consideration is **confidence level** (also called **reliability)** - i.e. how confident do we need to be in the answer. We generally work to a confidence level of between 80% and 95%, although 95% is only used in consumer marketing as sample sizes need to be larger, thus costs increase. In other words, what we have experienced in the test is likely to happen when we roll out to the larger list 8 times out of 10 (80% confidence level), 9 times out of 10 (90% confidence level) or 19 times out of 20 (95% confidence level).

**Note the use of the phrase ‘likely to happen’** – even after our test we are only able to say what is **probable –** thus, confidence level is sometimes also referred to as **probability** or **significance level** e.g. 80% or 95% probability or significance.

Next, we have to consider **error tolerance.** This is the allowable error – a plus or minus amount we must allow when reading our test results. In other words, again because of the imprecision of statistical laws, a 2% response to a test cannot be taken as exactly 2%, but as 2% + or – an amount which is determined by the number of responses we have received. This of course is a simple function of the number of names we have mailed multiplied by the response rate.

**What this means in general, is that the more names we mail in our test sample, the more we can rely on the result.**

Now before we look at a few examples we need to consider the formula for calculating sample size, error tolerance and reliability of results.

We will start with sample size. Firstly, let’s dispel a couple of myths. It is not unusual to hear experienced marketers say things like:

“You should always test 10% of the list”

“I never use test cells of fewer than 15,000 names”

The first statement is clearly nonsense; the second can often be wasteful – though as we shall see, such an approach may sometimes be prudent.

There is a formula for calculating sample size, or if we know the sample sizes and the response rate, for determining the error tolerance we need to apply for any given degree of reliability.

**It thus enables us to:**

1. **Decide on an appropriate sample size for our test**
2. **Evaluate a result in retrospect and thus decide how reliable a test is**
3. **Predict the range of response we can expect if we ‘rollout’ the test to the larger list of which our test sample is truly representative.**

To use the formula for calculating sample size you need to have an idea of the likely response you are expecting and the degree of error allowance you are prepared to tolerate. Whilst this may seem onerous, you would not really wish to embark on a mailing, even a test mailing, without some idea of what sort of response you expected.

The formula is:

# Sample size = (confidence level)2 x expected response x non-response

 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

 (error tolerance)2

Confidence level is expressed in standard deviations e.g.

 95% confidence = 1.96 standard deviations

 90% confidence = 1.645

 80% confidence – 1.281

Expected response and non-response are simple percentages e.g. 2.5% - expressed as 2.5; 97.5% expressed as 97.5

Error tolerance is the +/- amount you are prepared to accept – e.g. 0.5% (0.5)

We must remember that the formula gives us error tolerance2 so having arrived at this figure we hit the square root button to find the final error tolerance figure.

In marketing, we normally work to either 80% or 90% confidence though statisticians like to be more precise preferring 95% or even 99% confidence – they don’t have to pay for the names of course!

**NOTE** – you don’t need to memorise these formulae. But please work through the examples to get a feel for the process. Later, you’ll be able to use the Excel tool, based on the same formulae, that will do all the calculations for you.

## Now we will demonstrate how this works.

Let’s take a sample size of 5,000 names and a response rate of 2% - what does the formula tell us about this result? Well if we want to be confident that the result will be repeated 9 times out of 10 (90% confidence level) the formula gives us:

 **5,000 = (1.645 x 1.645) x 2 x 98**

 **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

 **(error tolerance)2**

1.645 x 1.645 x 2 x 98 = 530.3809

Thus, reversing our equation:

(error tolerance)2 = 530.3809 / 5,000 = 0.10607618

The square root of 0.10607618 is 0.3256933834145238

Thus, our error tolerance in this case is 0.33%

**This means our test that produced a 2% response, can be relied upon to deliver somewhere between 1.67% and 2.33% when ‘rolled-out’ - 90% of the time.**

If we wanted to have 95% confidence then the calculation would be:

 **ET2 = (1.96 x 1.96) x 2 x 98**

 **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

 **5,000**

Thus: ET2 is 0.1505907 so ET is 0.39 (0.3880601)

**As you can see the requirement for greater confidence increases the possible range of responses on roll-out to: 1.61% at worst to 2.39% at best – but we can feel confident that our roll out will be within this wider range 19 times out of 20.**

So far so good – if 1.61% is sufficient to give us an acceptable business result we can confidently proceed with the roll out.

Now let’s use the same formula to select an appropriate sample size

Now we estimate the **likely** response we are expecting and the degree of error allowance we are prepared to tolerate. So here we say ‘expected response’ as we are guessing in this case.

To remind you, the formula is:

# Sample size = (confidence level)2 x expected response x non-response

 **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

 (error tolerance)2**

So, allowing for **90% probability**, if we expect a **response of 2.5%** and we want to keep the **error tolerance** small, let’s say no more than **0.25%,** we see the following:

(1.645 x 1.645) x 2.5 x 97.5
 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

 (0.25 x 0.25) i.e. 0.0625

This gives: 659.5935937 divided by 0.0625 = 10553.49749 – a required sample size of say, 10,600.

Clearly, with such stringent requirements, if we want to test several things we will soon run up a very large bill.

What can we do to reduce the sample sizes and thus the cost?

We can consider our significance level. We have worked to 90% (9 times out of 10) – what would happen if we were prepared to accept an 80% probability level? Let’s work out the numbers:

**80% Significance Level**

 1.281 x 1.281 x 2.5 x 97.5

 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

 0.25 x 0.25 i.e. 0.0625

## This reduces the sample size to 6,400 (6,399.75)

What else could we do? Let’s say we are prepared to go to a 0.5% error tolerance – what difference does this make to the sample size?

 1.281 x 1.281 x 2.5 x 97.5

 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

 0.5 x 0.5 i.e. 0.25

**That’s better; our sample size now reduces to a mere 1,600 (1,599.93) Bear in mind here of course that all we would be able to predict from a 2.5% response from a sample of 1,600 names, is that the roll-out will produce somewhere between 2.0% and 3.0% - 8 times out of 10.**

## The ‘law of averages’ Some marketers (though never experienced statisticians) argue that, allowing for the ‘law of averages’, an 80% confidence level could be taken as 90% given the following scenario:

**80% confidence means that during ‘roll-out’:**

8 out of 10 will come in within the range defined by the +/- error limit.

1 out of 10 will produce less response

1 out of 10 will produce more response

**In fact, there is no ‘law of averages’ – the ‘law of the inertia of large numbers’ means that the above is possible but by no means a statistical probability, and we would need to carry out very many tests to expect an even spread above and below the expected range.**

However, given that 8 out of 10 probability is not bad odds and it is at least likely that **some** of the outliers will be above the range, 8 out of 10 could reasonably be considered as 8+. So, if budgets are tight, or available numbers are small, you might be prepared to accept 80% confidence for some of your test programmes.

**Had enough yet? Sorry we have hardly started. All we have done so far is to show how confident we can be in evaluating or predicting a roll-out from a single test cell.**

In many cases (probably most cases) we will want to compare a result against a control. And when comparing results with each other we use a different formula.

Here we may want to see whether say, a 2.5% test result is significantly better than a 2.0% control. On the face of it the test has out-performed the control by 25% - sounds great. But it really does depend on whether the samples are large enough to allow the result to stand once we have taken allowable error into account.

In this case, we use a formula that predicts the **probable difference** between two results according to the expected variation brought about by the imprecision of the laws of statistics. The formula predicts a ‘normal’ variation (or spread of responses) and we then see whether the gap between our two test cells exceeds this – if so the result can be relied upon – if not – no result! This formula is as follows:

**Expected difference = Confidence level x √ r1 (100 - r1) + r2 (100 - r2)**

 **n1 n2**

Where – confidence level is expressed as standard deviations (as before)

r1 = response to test cell 1 (% in decimal format)
r2 = response to test cell 2 (% in decimal format)
n1 = number mailed in test cell 1
n2 = number mailed in test cell 2

**Now let’s try this with the above test result, using samples of 5,000 per cell at a 90% confidence level.**

Sample 1 gets 2% response

Sample 2 gets 2.5% response. The expected difference is

**1.645 x √ 2 x 98 + 2.5 x 97.5
 5,000 5,000**

So: 1.645 x square root of: (0.0392 + 0.04875)

i.e. 1.645 x square root of 0.08795 which is 0.2965636

**Expected difference is: 1.645 x 0.2965636 = 0.487847122 - 0.49**

**Observed difference (what actually happened) is 0.5 (2.0% compared to 2.5%)**

**The result is significant –** **but only just.**

**Here’s another quick example:**

Sample 1 gives 1% response; Sample 2 gives 1.2%

Sample size was 7,000 for each cell and again we are looking for 90% confidence

**So: expected difference is 1.645 x square root of

 (1 x 99 / 7,000) + (1.2 x 98.8 / 7,000)**

 **(0.0141428) + (0.0169371) = 0.0310799**

Square root of this is 0.1762949 x 1.645 = 0.29000511

**So: expected difference is 0.29 Observed (actual) difference is 0.2 - This result is not significant**

## Alternatives to working with formulae

The easiest way to use these formulae is to input them into Excel. However, to give a simple demonstration, the following tables are included:

**NOTE – the tables are constructed using the same formula but if you work them out yourself you will find some small differences in the numbers of sample sizes. This is simply because of rounding and does not affect their usability.**

**This first table is based on 90% confidence level** and shows required minimum sample sizes for range of responses from 0.5% to 10%, and a range of +/-error limits from 0.1% to 1%.

**Table 1 Error tolerances @ 90% confidence level**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Response %** | **Error tolerances @ 90% confidence level** |  |  |  |
|  | **0.1%** | **0.25%** | **0.50%** | **0.75%** | **1%** |  |  |  |
| **0.5%** | 13462 | 2154 |  |  |  |  |  |  |
| **1.0%** | 26790 | 4286 |  |  |  |  |  |  |
| **1.5%** | 39982 | 6397 | 1599 |  |  |  |  |  |
| **2.0%** | 53038 | 8486 | 2122 |  |  |  |  |  |
| **2.5%** | 65959 | 10553 | 2638 | 1173 | 660 |  |  |  |
| **3.0%** | 78745 | 12599 | 3150 | 1400 | 787 |  |  |  |
| **4.0%** | 103911 | 16626 | 4156 | 1847 | 1039 |  |  |  |
| **5.0%** | 128536 | 20566 | 5141 | 2285 | 1285 |  |  |  |
| **7.5%** | 187730 | 30037 | 7509 | 3337 | 1877 |  |  |  |
| **10.0%** | 243542 | 38967 | 9742 | 4330 | 2435 |  |  |  |

**Note that minimum sample size increases when we want smaller tolerances.** So, if we expect a 5% response and are prepared to accept a result of 5% + or – 0.5% (i.e. our test can only be relied upon to produce between 4.5% and 5.5% when re-mailed) a sample size of 5,141 will be adequate. On the other hand, if we want the error limit to be halved – to 0.25% the minimum sample size must be 4 times as large (20,566).

**Table 2 Error tolerances @ 80% confidence level**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Response %** | **Error tolerances @ 80% confidence level** |  |  |  |
|  | **0.10%** | **0.25%** | **0.50%** | **0.75%** | **1%** |  |  |  |
| **0.5%** | 8151 | 1304 |  |  |  |  |  |  |
| **1.0%** | 16220 | 2595 |  |  |  |  |  |  |
| **1.5%** | 24207 | 3873 | 968 |  |  |  |  |  |
| **2.0%** | 32113 | 5138 | 1285 |  |  |  |  |  |
| **2.5%** | 39936 | 6390 | 1597 |  |  |  |  |  |
| **3.0%** | 47677 | 7628 | 1907 | 848 |  |  |  |  |
| **4.0%** | 62915 | 10066 | 2517 | 1118 |  |  |  |  |
| **5.0%** | 77824 | 12452 | 3113 | 1384 |  |  |  |  |
| **7.5%** | 113664 | 18186 | 4547 | 2021 | 1137 |  |  |  |
| **10.0%** | 147456 | 23593 | 5898 | 2621 | 1475 |  |  |  |

As you can see, being prepared to accept a lower confidence level makes required minimum sample sizes much smaller.

Let’s look at a couple of examples to see how these tables can be used. For simplicity, I’ll work these examples using only the 80% confidence table.

Let’s take a sample size of 5,000 names and a response rate of 2% - what can we predict from this? Looking this up in the 80% confidence table we see that a response rate of 2% from a sample of 5,138 (the nearest number quoted to our sample size) means we must allow for error tolerance of 0.25%. **This means that our test that produced a 2% response, can only be relied upon to deliver somewhere between 1.75% and 2.25% when repeated (or ‘rolled-out’).**

As we are using the 80% table we can say that when rolled out to the larger list (of which our sample was truly representative) we can expect a response rate of between 1.75% to 2.25%, **8 times out of 10.**

So far so good – if 1.75% is sufficient to give us an acceptable business result we can confidently proceed with the roll out.

Now let’s put some of this into practice with a few examples. Here’s a test matrix we might construct:

**Table 3 A possible test matrix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Control****Offer** | Offer 2 | Offer 3 | Total |
| Control list | **3,500** | **3,500** | **3,500** | **10,500** |
| List B | **3,500** | **3,500** | **3,500** | **10,500** |
| List C | **3,500** | **3,500** | **3,500** | **10,500** |
| Total | **10,500** | **10,500** | **10,500** | **31,500** |

## What can we learn from such a test?

As we have seen, the sample sizes used here would enable us to compare list with list (across all offers) and offer with offer (across all lists) with a fair degree of confidence. Our total sample for each of these is 10,500 and at 2.5% response this gives us an error tolerance of 0.25% at 90% confidence level, or 0.2% at 80% confidence level.

Of course, what we would really like to do is to compare what happened in each of the individual cells and if we do this we will find that the error tolerance will change because the sample sizes are smaller. A response rate of 2.5% to a sample of 3,500 requires an error tolerance of +/- 0.43% at 90% or 0.34% at 80% confidence level. This is not a problem but, as we have seen, it means that differences must be larger to be significant.

## Can we reduce test costs still further?

If we are prepared to take a slightly increased chance of our results not being totally reliable the easiest thing is to use a lower confidence (or significance) level. Let’s look at another text matrix, this time using smaller samples:

**Table 4 Testing with smaller samples**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Control****Offer** | Offer 2 | Offer 3 | Total |
| Control list | **1,300** | **1,300** | **1,300** | **3,900** |
| List B | **1.300** | **1.300** | **1.300** | **3,900** |
| List C | **1.300** | **1.300** | **1.300** | **3,900** |
| Total | **3.900** | **3.900** | **3.900** | **11,700** |

As before this will probably tell us which list and which offer were more successful overall (2.5% response from a sample size of 3,900 gives error tolerance of 0.41% at 90% and 0.32% at 80% confidence level). Trying to compare individual cells at this level is not recommended. The error tolerance for samples of 1,300 at 2.5% is huge (0.71% at 90% confidence, and 0.55% at 80%) making reliable comparisons difficult.

## A few final comments

1. Test sample sizes are a matter for statistical calculation rather than guess work.
2. We can use the formula to select a suitable sample size and, in reverse, to assess the degree of reliability of any given test.
3. It is often easier to find a friendly marketing statistician than spend too much time learning and practising with formulae, although using the Excel tool makes the process very easy.

**Statistical significance with very small response rates or uncertain sample sizes**

As mentioned earlier, the above formulae for confidence level and error tolerance based on sample sizes do not work when we do not know the exact sample size nor for any tests where the response rate is lower than 0.1%. So, it is not used to evaluate tests in press advertising, loose inserts, TV commercials or banner ads.

To evaluate a newspaper split run test for example we use the chi-squared test. To apply this test, we add up the total responses to both halves of the test (A + B). Then we take the larger of the two response numbers (A or B) and calculate what percentage that is of the total.

**Table 5 Significance Factors in Split Run Testing - 90% Confidence Level**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Total Response****i.e. A + B** | **Significance Factor** | **Winner** | **Loser** | **% Gain** |
| **100** | **60%** | **60** | **40** | **50%** |
| **200** | **57%** | **114** | **86** | **33%** |
| **400** | **55%** | **220** | **180** | **22%** |
| **600** | **54%** | **324** | **276** | **17%** |
| **1,000** | **53%** | **530** | **470** | **13%** |

First of all, we can see that the greater the number of replies in total, the closer the difference can be between winner and loser. The numbers in column 2 are the percentage of the total response that the more successful half of the split run must reach before we can be confident that the same ad would win 9 times out of 10.

With a total response of 400, to be sure we have a ‘winner’ 9 times out of 10 the winning ad must have achieved at least 220 replies (55% of 400). These are not small differences – if the winner got 220 of the 400 replies the loser only received 180. **So, the winner shows a gain of 40 over 180 – a percentage increase of 22.2%.**

**With a total of only 100 responses the winner must receive at least 60 against 40 – an increase of 50%.**

## A final word of caution

The numbers in Table 5 are the absolute minimum that a ‘winner’ must receive. It would be prudent to re-run anysplit run testwhere a marginal result occurred. A sensible rule of thumb might be that unless you have a large number of responses, say at least 600, you should disregard (or at least repeat) any split run tests where the difference between winner and loser is less than 25%.

## Reliability of newspaper split runs

Some critics say A/B splits are unreliable, citing a case where by mistake the same advertisement was run in both halves of the split. One half of the test produced 15% more response than the other but this does not prove unreliability; it simply demonstrates that the commentator does not understand statistics.

Of course, there will be a variation; that is why we calculate the **expected variation,** and thensee whether our ‘winner’ has produced a response that exceeds this or not.

## Don’t stop monitoring

Measurement of response is only the start of course. You must continue to monitor the ongoing behaviour of respondents to ensure that an apparently successful new idea is not simply producing a large volume of poorly qualified enquiries.