

The lady tasting tea

Using experimental methods to introduce inference statistics

Bjørn Felsager

Mathematics teacher, Emeritus, Midtsjællands Gymnasieskoler, Denmark,

Bjoern.Felsager@Skolekom.dk

1. Introduction

In the upper secondary school in Denmark inference statistics has recently (august 2010) changed its status from being a voluntary subject to being a compulsory subject. Danish students can choose among three levels of mathematics, a one-year C-level, a two-year B-level and a 3-year A-level. All students in the STX gymnasium (General studies) and the HHX-gymnasium (Mercantile studies) having math at least at the B-level are required to be able to handle inference statistics in the form of the χ^2 -tests of Goodness of Fit as well as of Independence. The goal is to strengthen the collaboration with other subjects using inference statistics, in particular the Social Sciences. A formal background in probability theory is not mandatory, so it has been necessary to develop an experimental approach relying on statistical concepts rather than formal probability theory. In particular the Danish ministry of education has emphasized that a formal treatment of stochastic variables is not required.


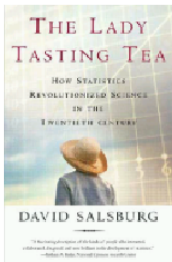
In the following I will outline one possible approach, but there are actual several possible approaches and teachers are free to teach χ^2 -tests any way they like. I will take as a starting point the fundamental concept of null hypothesis and demonstrate how to simulate a null hypothesis. This not only gives a very good feeling for what a null hypothesis actually is but it also makes it possible to actually build up the distribution of a test statistics and thus explore the meaning of significant events. I will begin with a famous and very illustrative historical example, The lady tasting tea, which is in fact closely related to the χ^2 -tests of independence, and then proceed to discuss the machinery involved in independence tests.

2. Lady Tasting Tea

Ronald Fisher developed basic methodology of modern inference statistics while he was working at the Rothamsted Plantation Station in the 1920's. There he introduced basic concepts such as *the null hypothesis* and *the level of significance*. He collected his methodology in the highly influential book *Statistical Methods for Research Workers* from 1925, which contained a separate chapter about the Design of Experiments. This chapter gradually expanded into its own highly influential book, *The Design of Experiments*, from 1935. In the opening chapter Fisher explains his basic thought using a now very famous experiment concerning the *Lady Tasting Tea*. Recent investigations, in particular interviews with his daughter, suggest that the experiment was actually performed in the 1920's although details may vary depending on who you ask about the events, which at the time of the investigation took place at least 50 years ago!

So here is the popular version of what happened: At a summer day at Rothamsted Plantation Station after work Fisher wanted to be gallant to a young biologist Muriel Bristol and offered her a cup of tea. But when she asked him if he had remembered to pour in milk first he confessed that he had not, but also argued that he could not see the point: Once the milk and the tea is mixed surely it is impossible to taste the difference. But Muriel Bristol claimed that notwithstanding his objections she was actually able to taste the difference; and so she went to pour her own cup of tea. The fiancée of Muriel Bristol standing in the background of the room had observed the incident and shouted to Fisher: "You can test her!" And that's what Fisher did, thus immortalizing the Lady tasting tea as the archetypical statistical experiment. The Lady Tasting Tea is also the title of a very enjoyable book by David Salsburg about the modern history of statistics. Fishers chapter is very informative and you can actually use it in your own teaching: It covers all the basic concepts of inference statistics with very clear and lucid explanations and arguments!

The Lady Tasting Tea



Fishers own introduction to tests of significance from [The Design of Experiments \(1935\)](#)

A LADY declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup: We will consider the problem of designing an experiment by means of which this assertion can be tested. For this purpose let us first lay down a simple form of experiment with a view to studying its limitations and its characteristics, both those which appear to be essential to the experimental method, when well developed, and those which are not essential but auxiliary.

Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order. The subject has been told in advance of what the test will consist, namely that she will be asked to taste eight cups, that these shall be four of each kind, and that they shall be presented to her in a random order, that is in an order not determined arbitrarily by human choice, but by the actual manipulation of the physical apparatus used in games of chance, cards, dice, roulettes, etc., or, more expeditiously, from a published collection of random sampling numbers purporting to give the actual results of such manipulation. Her task is to divide the 8 cups into two sets of 4, agreeing, if possible, with the treatments received.

In the above slide there is an outline of the experiment involving eight cups of tea, four of which are prepared with milk first whereas the remaining four are prepared with tea first. Notice also the emphasis on randomness and the description of how it is obtained, either by using a physical apparatus, such as throwing a dice, or by looking up in a table of random numbers. Today the table is replaced by a computer, but essentially the computer does the same: It has built in random function that is able to compute a very long string of random numbers.

On the next slide you will find Fishers characterization of the null hypothesis and the level of significance:



Fisher: This hypothesis, which may or may not be impugned by the result of an experiment, is again characteristic of all experimentation. Much confusion would often be avoided if it were explicitly formulated when the experiment is designed. In relation to any experiment we may speak of this hypothesis as the 'null hypothesis,' and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.

Null hypothesis: Pure randomness! What we observe comes about exclusively as a consequence of random variations.

Fisher: It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result. ... It is usual and convenient for experimenters to take 5 percent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard

Level of significance: The traditional level is 5%. If the probability of a random event, that is at least as extreme as the observed event, falls below 5%, the null hypothesis is rejected as a reasonable explanation of what we have observed.

Notice in particular *the null hypothesis* always explains the observed outcome as a result of pure randomness. So in the case of the Lady Tasting Tea the null hypothesis claims that she is merely guessing and that the outcome of the experiment is the result of pure chance: For every cup she thus has the probability $\frac{1}{2}$ of guessing whether the milk was poured in first or last. This of course is precisely the opposite of Muriel's claim: That she can actually taste the difference and that she is certainly not guessing anything. The point however is, that you can draw consequences of randomness and thus you can investigate the Null hypothesis in a pure mathematical fashion. Depending upon your analysis you can then make a rational objective choice between the null hypothesis of pure randomness and the alternative hypothesis: That she can to some extent actually taste the difference.

Notice also Fishers insistence upon the important fact that you can never prove or establish a null hypothesis, but you can possibly disprove it.

At this point we want to build a simulation of the null hypothesis and there opens a List and Spreadsheet application in TI-Nspire™ CAS. We enter a list called **tray** with eight values: Four "Milk" and four "Tea" corresponding to the physical setup used for the preparation of the cup of tea, pouring in milk first or tea first. Notice the quotation marks used for entering "Milk" etc. The variable **tray** is a categorical variable and has text-strings as its values. Unlike Excel it is now important to enclose the values in quotation marks, because TI-Nspire™ CAS is a symbolic spreadsheet and without quotation marks it will interpret the text as a mathematical expression which may cause some confusion☺.

	A	B
◆		
1	"Milk"	
2	Milk	
3	Milk	
4	Milk	
5	Tea	
6	Tea	
7	Tea	
8	Tea	

So this is a model of the tray that is going to be tested. But according to the null hypothesis the identification of the cups happens by pure chance, so all we have got to do is to make another random list containing the same 8 cups in a completely random order. This is done using a RandSamp command for making a random sample from the tray. The RandSamp-command has three arguments as you can see from the catalogue:

```
randSamp(List, #Trials [,noRepl])
noRepl=0 with replacement (default)
noRepl=1 without replacement
```

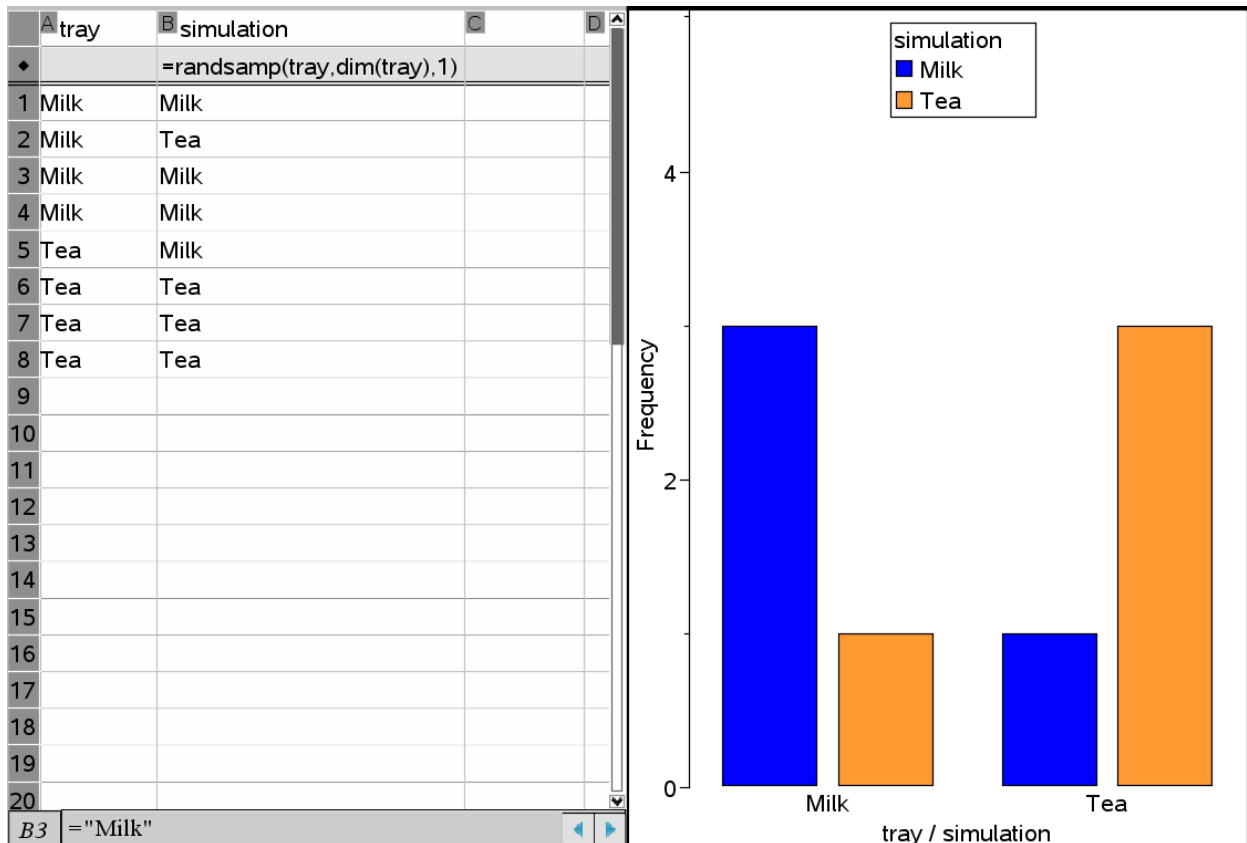
	A tray	B simulation
◆		=randsamp(tray,dim(tray),1)
1	Milk	Tea
2	Milk	Milk
3	Milk	Tea
4	Milk	Tea
5	Tea	Milk
6	Tea	Tea
7	Tea	Milk
8	Tea	Milk

You have to specify a list to draw the sample from (the population), the number of elements you sample and optionally whether the sampling is with replacement (default) or without. In our case sampling must be without replacement, so we really need the additional parameter 1. The sample is called **simulation**, because it reflects the null hypothesis of guessing the type of the first cup, the type of second cup etc. Notice that the dim(**tray**)-command computes the number of elements in the list **tray**. This allows you to extend the experiment by adding further cups.

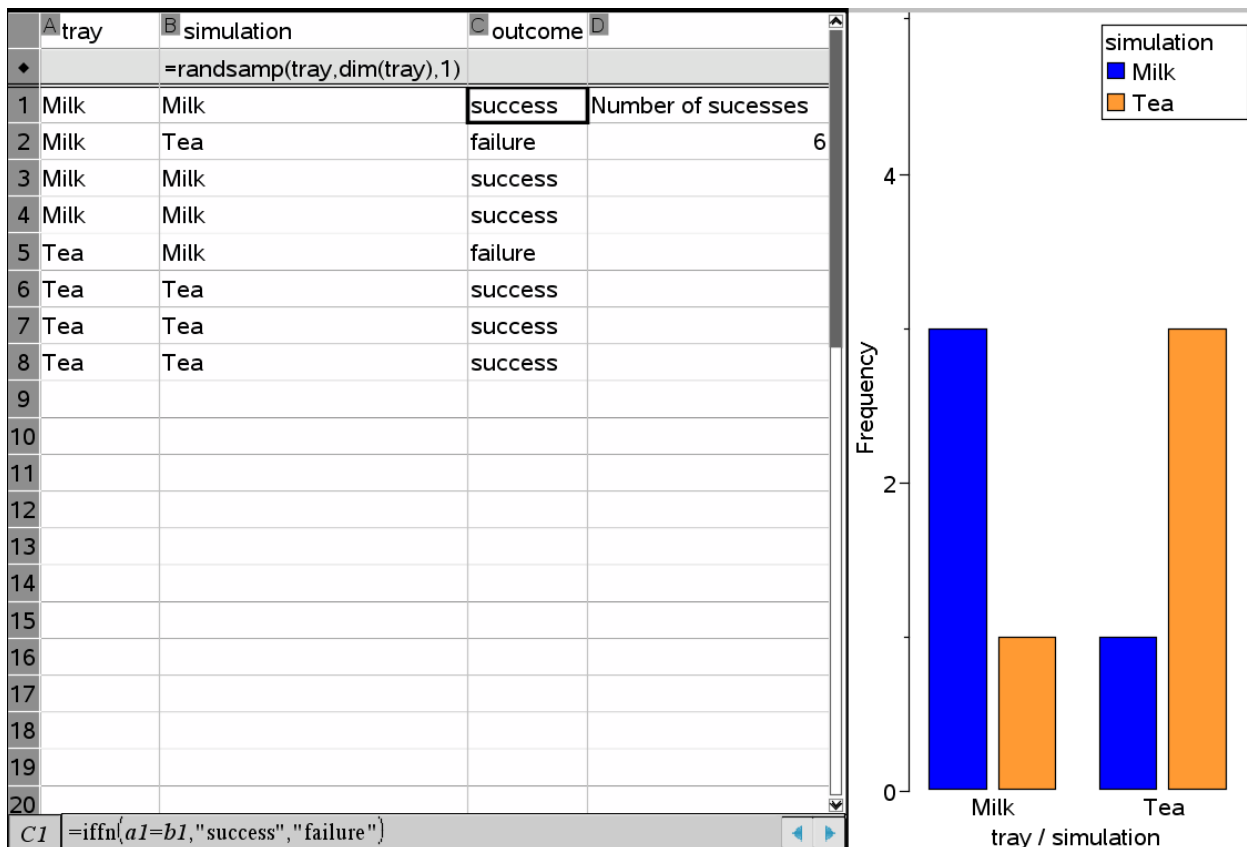
In the above simulation we got the second and the sixth guess correct, so we got a total of 2 successes.

Remark: When doing this with the class you will observe that all students in the class obtain identical random simulation! This is somewhat disturbing to the students: How can they get identical results, when they are supposed to be completely random? The short answer is that the simulation is built up looking up random numbers in the built in 'table' of TI-Nspire™ CAS. And if you do nothing to prevent it everybody is going to look up random numbers from the very beginning of the table. But we can easily repeat the simulation by pressing **CTRL R** for recalculation. If you keep pressing down **CTRL R** for some while you will progress to different parts of the table and your results will no longer be in sync!

We can also construct a diagram reflecting the outcome. To do that, we split the window and open a Data and Statistics application. We map the variable **tray** along the first axis and then *right click on the first axis* to split the categories of the variable **tray** with the variable **simulation**. The resulting dot chart can easily be converted to a bar chart, where the second axis is adjusted to show a maximum of four cups of a given type. In the milk section the number of blue milk-guesses are successes, where as in the tea section the number of orange tea-guesses are successes. By construction of the simulation these two numbers always coincide, so we only have to observe the number of blue milk-guesses and double up!



It is however also easy to actually count your successes. To do that we add a column called outcome and we use the conditioned cell-command `ifn()` to find out whether the two neighbouring cells are identical or not.



`C1 =ifn(a1=b1,"success","failure")`

Once you have inserted the cell-command you can fill it down along the list or you can simply drag it down using the anchor in the lower right corner!

We have also counted the number of successes in the next column using the cell-command:

```
D2 =countif(outcome,"success")
```

Notice that this column has no name! Thus it is not a part of the list-component of the spreadsheet and the cells can be used freely in the same way you use cells from Excel!

Coming that far we can make a proto-test of the significance of the Lady Tasting Tea! We need two facts:

First of all we need an observed event, i.e. we need to know what actually happened the afternoon the experiment was performed: How many successes did Muriel score in the experiment. It turns out she made a *'full house'*, i.e. 8 successes.

Next we need to agree on a level of significance, which should be agreed upon before the experiment is performed! We will use the standard level, i.e. 5%. This corresponds to the ratio 1:20. We will therefore conduct 20 simulations and see if we can replicate a single success.

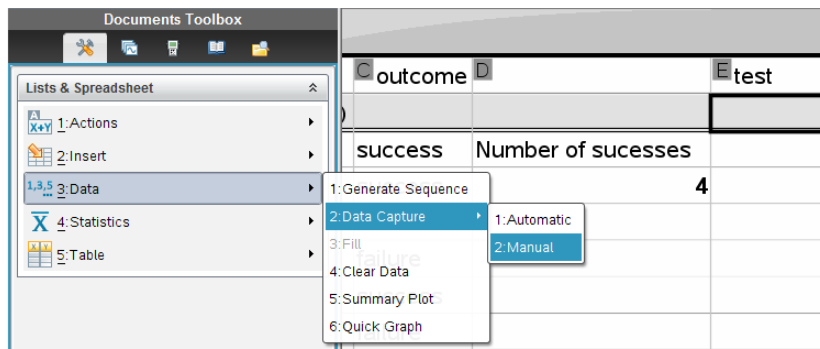
Pressing **CTRL R** 20 times will do the job. In my case e.g. I got no full houses. This means that the proportion of extreme events, events that are at least as extreme as the observed event, seems to fall below 5%. This indicates strongly that simulating a *'full house'* is not easy and therefore that the null hypothesis is not a convincing explanation for the observed event. At this point we therefore seem to have reasons for rejecting the null hypothesis and thus declaring the observed event statistically significant. Muriel is vindicated! To believe her claim seems more reasonable than to doubt it.

Still you might be worried that 20 simulations is a little shaky ground for making your decision! Standard in industrial tests of significance is 500 simulations. So clearly we need a procedure for automatizing the simulations. This can be done using data capture. This is slightly more technical than the bare simulation of the null hypothesis, so we will now explain the necessary steps in some details.

First we need to be able to capture the number of successes. This value of the cell must therefore be stored in a variable, which we will call **test_sim**. The easiest way to do this is to put the variable name followed immediately by a colon in front of the equality-sign:

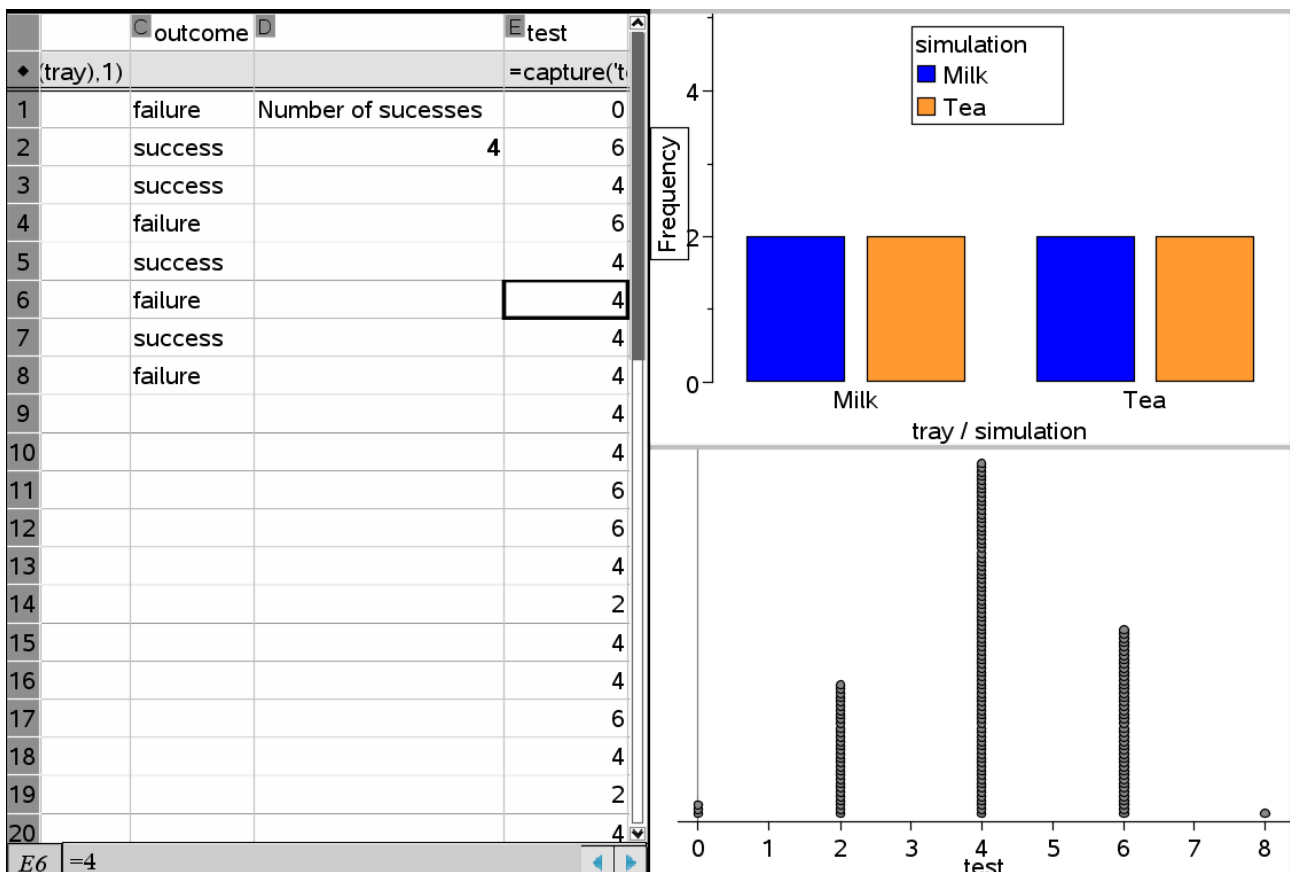
```
D2 test_sim:=countif(outcome,"success")
```

The cell is now displayed in bold to emphasize that it now reflects the value of a captured variable. Next we introduce a list named **test** in the next column, press Enter to move to the formula cell right below the list name. But we don't enter this field, we only select it – so don't hit Enter twice! We can then go to the **Data** menu and select the command **Data capture manually**:



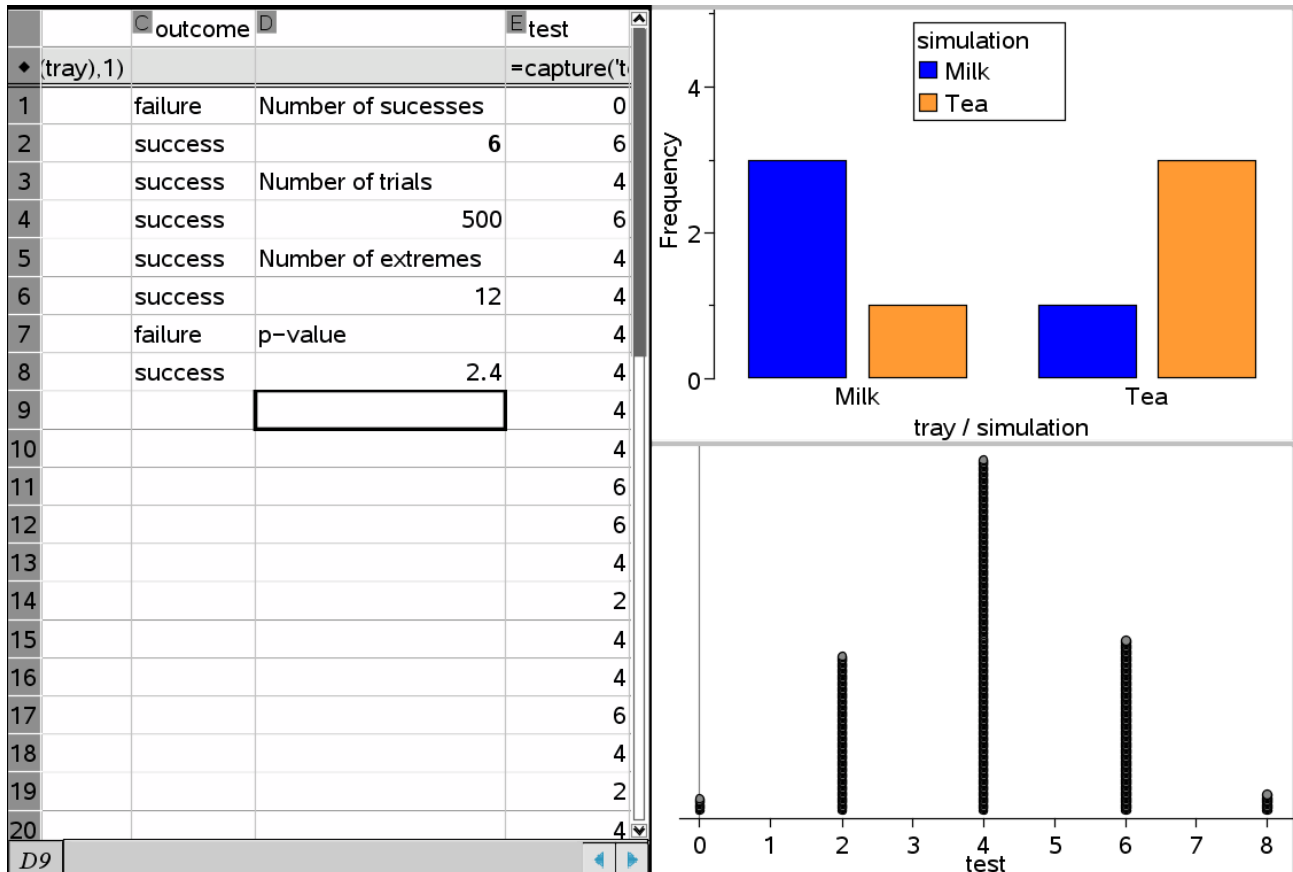
You may wonder about the difference between an automatic and a manual data capture: Automatic Data Capture only captures the value of the variable when it changes! But it does it automatically every time it changes. Manual data capture only captures the value when it is instructed. But then it does whether or not the value has changed. So manual data capture is safer. It requires however a capture command, which turns out to be CTRL . (i.e. Control Dot), so the variable is captured every time you hit CTRL .

Next you specify the name of the variable you want to capture, in this case **test_sim**. And you are ready to start hunting. To see the effect of the hunting we also split the window further to produce another diagram in the Data and Statistics Application. This time we graph **test** along the first axis and adjust the axis to show values from 0 to 8. Make sure the List and Spreadsheet application is activated! Now keep pressing the **CTRL** key with one finger and alternate between **R** and **DOT** using two other fingers. You then start seeing the distribution of the number of successes slowly being built up:



We continue until at least all possible events have occurred. We can count the number of simulations as well as the number of extremes and their proportion (the p-value) using the cell commands:

D4	=dim(test)
D6	=countif(test,8)
D8	$=\frac{d6}{d4} \cdot 100.$



Thus we see that the estimated p-value is 2.4 %, which falls short of the level of significance 5%, so the null hypothesis is rejected and Muriel’s claim is vindicated once again!

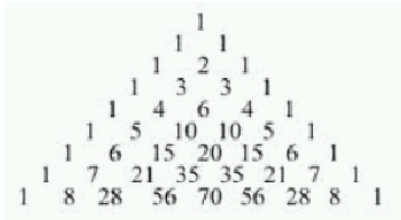
At this point we have solved the problem experimentally by simulating the null hypothesis. In many cases it is possible to do a theoretical analysis of the problem and actually compute the expected p-value. Fisher notices the possibility in the present case using only very elementary combinatorics. To make the analysis slightly more adaptable to variations of the test using e.g. a different number of cups, I have included a Pascal Triangle to keep count of the number of cases for the different outcomes:

The tray holds four cups made with milk first and another four cups made with tea first.

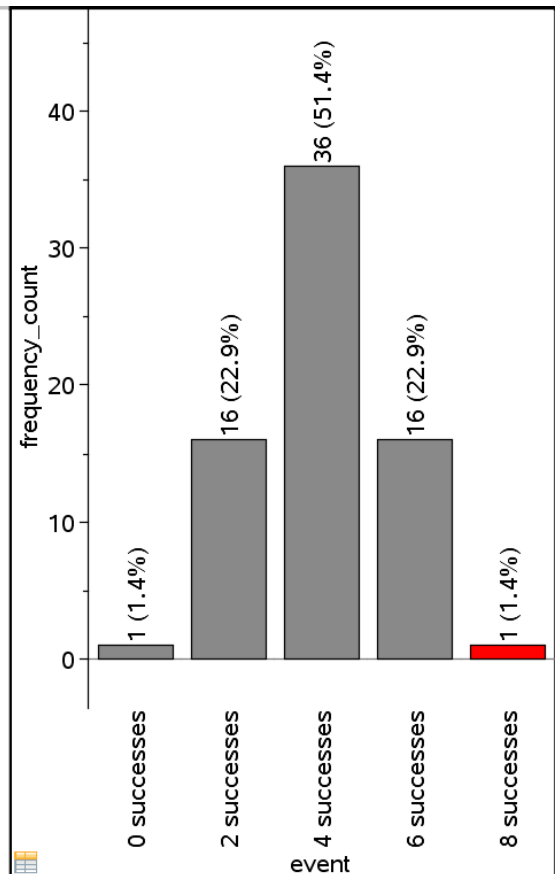
Tray: Milk, Milk, Milk, Milk Tea , Tea , Tea , Tea

To make a guess we must distribute four guesses with Milk and another four guesses with Tea. To make a score of 8 successes: There is exactly one distribution that full fills this demand: Four milks combined with four teas.

Theoretical analysis: Counting events



- Tray:** Milk, Milk, Milk, Milk Tea, Tea, Tea, Tea
- 8 succ.:** Milk, Milk, Milk, Milk Tea, Tea, Tea, Tea
- 1x1**
- 6 succ.:** Tea, Milk, Milk, Milk ...
- 4x4** Milk, Tea, Milk, Milk
Milk, Milk, Tea, Milk
Milk, Milk, Milk, Tea
- 4 succ.:** Tea, Tea, Milk, Milk
- 6x6** ...
- 2 succ.:** ...
- 4x4** ...
- 0 succ.:** ...
- 1x1** ...



To make a score of 6 successes: This means we fail one milk and similarly we fail one tea. So this time we have to combine one failure of the milks with one failure of the teas, which leads to 4x4 outcomes as the failure of milk can be distributed in exactly four positions etc.

Continuing this way we see that the expected frequencies of the different outcomes corresponds exactly to the squares of the numbers in the fourth row of the Pascal Triangle! Furthermore the total number of expected outcomes corresponds to the central number of the eighth row of the Pascal Triangle. Thus the expected p-value is 1/70 or 1.4%. We see that the estimate differs from the theoretically value with 1%. To obtain a more precise estimate one must therefore make more simulations!

The expected distribution has been diagrammed as a summary chart (notice the icon in the lower left boundary) based upon a manual entering of the cases and frequency counts at the end of the spreadsheet. To produce a summary chart you right click in the first axis:

Y event	Z frequency_count
8 successes	1
6 successes	16
4 successes	36
2 successes	16
0 successes	1

Clearly the above reasoning can be extended to cases of 10 cups or 12 cups on the tray.

Finally we summarize the discussion using the familiar metaphor of a significance test as being like a Trial at the Court:

The metaphor of a Trial at the Court

The Null Hypothesis is put on trial and cross-examined by the prosecutor, who tries to cast doubt on the credibility of the null hypothesis' claim: **The observed event can be explained as resulting from pure randomness.**



The null hypothesis is given the opportunity to justify his claim by means of repeated simulations to try to produce an event that is at least as extreme as the observed.

The result of the cross-examination is called the **p-value indicating the proportion of extreme outcomes.**



The judge balances the p-value off the level of significance:

If the level of significance level is heavier than the null hypothesis, the null hypothesis is too light and is rejected as being untrustworthy as an explanation for the observed event.

If on the contrary the null hypothesis is heavier than the level of significance, the null hypothesis is acquitted.

This concludes our discussion of the Lady Tasting Tea!

3. Significance test of independence

In the preceding discussion of the Lady Tasting Tea we were dealing with two 'identical' categorical variables **tray** and a classification of the cups according to Muriel's handling the test or a random simulation. Both variables had values consisting of four 'Milk' and four 'Tea' in a specific order. The main question was if these two variables, representing the preparation of the cups and the classification of the cups, were somehow related e.g. according to Muriel's claim that they were in fact identical or almost identical, or they were completely unrelated as was case with simulation of the null hypothesis.

In general we want to investigate the relationship between *two categorical values* that need not have any values in common. In the Social sciences e.g., one is interested in analysing questionnaires from surveys. These typical incorporate categorical variables, some of which are considered explanatory or independent variables, such as e.g. the variable **sex** with the values female and male; others are considered response variables or dependent variables, such as e.g. the variable **position**, whose values reflects the respondents agreement with a suitable

statement. Since we are dealing with categorical variables all we can do is to count the number of cases for all possible combinations of the values of the individual variables. This is recorded in a *summary table* (pivot table in Excel). A typical example may look like the following:

Simulating the null hypothesis in a test of independence

We take as a starting point a survey, where we want to examine the relationship between sex and political ideologies. In the survey the respondent must among other questions take a position with respect to the following statement:

To what extent do you agree with the following statement:
 "A certain measure of inequality is desirable, due to the difference in peoples effort."

I completely agree I almost agree I neither agree nor disagree I almost disagree I completely disagree

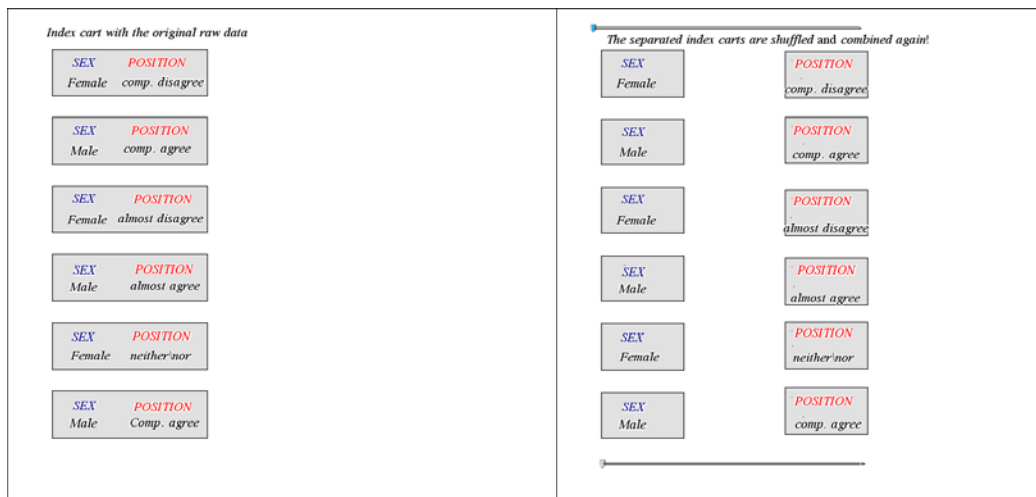
In a pilot survey the following frequencies were obtained:

	Female	Male
I completely agree	12	24
I almost agree	37	46
I neither agree nor disagree	22	20
I almost disagree	14	6
I completely disagree	9	5

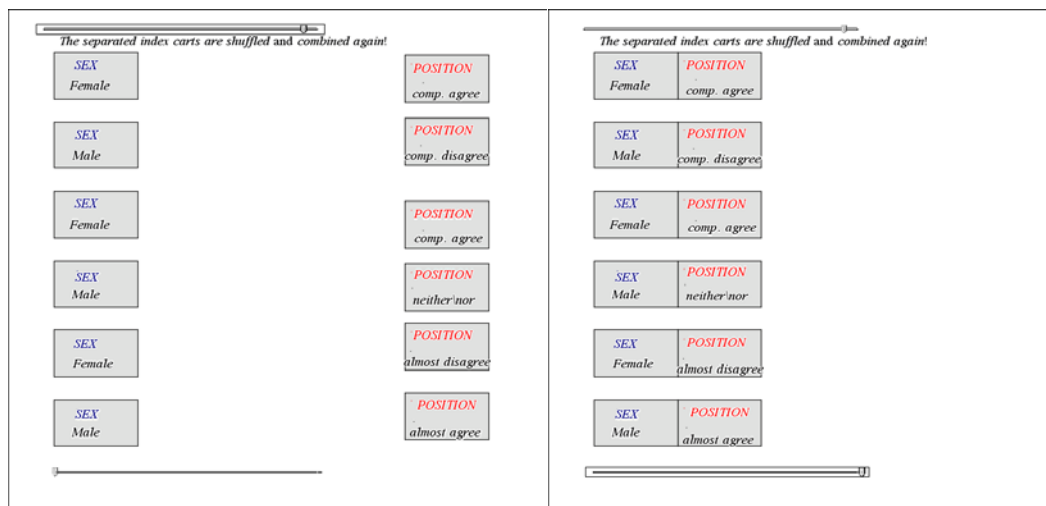
Investigate on the basis of these frequencies the following hypothesis:
 There is no difference between the position of men and women with respect to the above statement.

The question is whether there is a significant difference between the position of females and males, or on the contrary they follow the same distributions except for the inevitable random fluctuations resulting from the sampling. The latter hypothesis is thus the null hypothesis. But if the distributions are identical, we may also say that the variable **position** is independent of the variable **sex**. We thus have two equivalent descriptions of the null hypothesis: One is focusing upon the homogeneity of the samples (i.e. they follow identical distributions) the other is focusing upon the independence of the two variables (i.e. the distribution of **position** is independent of the **sex**, i.e. constant across the two sexes.) These two situations leads to the same test, sometimes called a test of homogeneity, at other times a test of independence. So there really is only one test. You may encounter a difference in the description of the sampling procedures for the two types of test: In the homogeneity test we focus upon two different populations: A female population and a male population, and we draw two distinct samples, one from each population. In the independence test, there is only one population and only one sample, which we split according to two categorical variables. But the procedures in both tests are identical, so in the following we will focus exclusively upon the independence test.

Next we want to *simulate the null hypothesis*, i.e. the independence of the two variables. This can be done using a clever procedure called *scrambling of the observed data*.



First we split the original index cards, so that the information about the respondent's **sex** is separated from the information about the respondent's **position**.



Then we perform an arbitrary permutation of the positions, i.e. we scramble the positions in a complete random manner. This scrambling clearly breaks any dependence that might have existed between the two variables. Finally we combined the original information of the sex with the scrambled information of the position, thus creating the simulation of the null hypothesis.

Notice that scrambling does preserve the so-called marginal totals of the variables involved. Although we scrambled the **positions** we did not alter the total number of respondents who agree completely with the statement etc. And since we did nothing to the sex-variable clearly we did not change the total number of females etc. We only changed the combination of the two variables!

Remark: Scrambling is the essence of the simulation of independence of two stochastic variables whether they are categorical or not. But the χ^2 -test will only apply to categorical variables. As soon as a numerical variable is involved one must invoke other types of tests, in particular t-tests.

We are now ready to analyse the observed summary table from our example. Notice how the summary table is entered 'below the bar', i.e. there are neither titles nor formulas in the

spreadsheet 'above the bar'. Everything is entered in Excel style – at this point no lists are involved!

	A	B	C	D
◆				
1	Observed	Female	Male	
2	comp. agree	12	24	
3	almost agree	37	46	
4	neither\nor	22	20	
5	almost disagree	14	6	
6	comp. disagree	9	5	
7				

On the assumption of the null hypothesis we first compute the expected values for the various combinations of the two variables. As a preliminary step this involves the calculation of totals, such as column-totals, row-totals and table-totals. This is done using the cell-command sum():

	A	B	C	D
◆				
1	Observed	Female	Male	Total
2	comp. agree	12	24	36
3	almost agree	37	46	83
4	neither\nor	22	20	42
5	almost disagree	14	6	20
6	comp. disagree	9	5	14
7	Total	94	101	195

D2 =sum(b2:c2)

Once you have calculated the total number in complete agreement with the statement, you fill down the cell-command (or drag it down along the table) to compute the remaining row-totals. Similarly you compute the total number of females using a sum()-command and drag it along horizontally to obtain the total number of males as well as the total number in the table!

According to the null hypothesis the female and male distributions are now ideally identical, which means they both are identical to the distributions of the row-totals i.e. the distribution of the positions in the combined sample including both female and male. According to the null-hypothesis we thus expect 36/195 of the female as well as of the males to agree completely. But since the total number of females is 94, this means that we expect $36/195 \cdot 94$ of the females to agree completely.

The actual calculation is done by first copying the original table of observed values and the dragging along cell-formulas built upon the above pattern. Care must be taken to differentiate between absolute references and relative references: Row 7 contains the column-totals and we must make sure that the row number seven is fixed during the calculation. Thus it should be preceded by a dollar-sign. Similarly column D contains the row-totals and we must make sure that column D is fixed during the calculation. Thus it must also be preceded by a dollar-sign.

	A	B	C	D
1	Observed	Female	Male	Total
2	comp. agree	12	24	36
3	almost agree	37	46	83
4	neither\nor	22	20	42
5	almost disagree...	14	6	20
6	comp. disagree	9	5	14
7	Total	94	101	195
8				
9	Expected			
10	comp. agree	17.35	18.65	36.
11	almost agree	40.01	42.99	83.
12	neither\nor	20.25	21.75	42.
13	almost disagree...	9.641	10.36	20.
14	comp. disagree	6.749	7.251	14.
15	Total	94.	101.	195.

$$B10 = \frac{d2}{d7} \cdot b7 \cdot 1.$$

We know that we have done it right when the column- totals and row-totals for the table of expected values sum up to the original values!

Remark: Notice that the expected values are decimal numbers, whereas the observed values were integers. This is okay because the expected values represent the mean values of an infinite number of random samples, and mean values may very well be fractional. We can now compare the similarity between the observed values and the expected values. According to the null hypothesis any differences are just due to random fluctuations in the observed sample. The difference between the observed values and the expected values are thus expected to be 'small'.

In 1900 Pearson made the crucial observation that you could measure the difference between the observed frequencies (the actual counts) and the expected frequencies using a suitable weighted sum of squared differences:

$$\chi^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}$$

This is used as the test-quantity for the χ^2 -test, with large test-quantities representing significant violations of the null hypothesis, thus forcing us to reject the null hypothesis of independence between the two categorical values.

In our case we can calculate this test-quantity directly from the observed and expected tables using that cell-ranges act like lists. Thus the values from the observed table is represented by the range b2:c6 and similarly the values from the expected table is represented by the range b10:c14. As long as you stay inside the spreadsheet you can perform list calculations directly without using named lists!

	A	B	C	D	E
1	Observed	Female	Male	Total	Chi2_obs
2	comp. agree	12	24	36	9.175
3	almost agree	37	46	83	Chi2_sim
4	neither\nor	22	20	42	4.273
5	almost disagree...	14	6	20	
6	comp. disagree	9	5	14	
7	Total	94	101	195	
8					
9	Expected				
10	comp. agree	17.35	18.65	36.	
11	almost agree	40.01	42.99	83.	
12	neither\nor	20.25	21.75	42.	
13	almost disagree...	9.641	10.36	20.	
14	comp. disagree	6.749	7.251	14.	
15	Total	94.	101.	195.	

$$E2 \quad \text{chi2_obs} := \text{sum} \left(\frac{(b2:c6 - b10:c14)^2}{b10:c14} \right)$$

In our case we thus obtain an observed χ^2 -value 9.175, which we have stored in the variable chi2_obs so that we can refer to it in diagrams etc. To decide whether this is a big value or a small value we must know its expected value according to the null hypothesis. Here the rule is very simple:

On the assumption of the null hypothesis the expected value of χ^2 is equal to the number of degrees of freedom.

In our case there are 4 degrees of freedom. This follows from the following observation: The survey includes a total of 94 females and 101 males. Similarly it includes a total of 36 individuals who completely agree, 83 individuals who almost agree etc. This put some restrictions on the outcomes. E.g. you can choose fairly freely the value the number of females that agree completely, say give it the value 20. But once you have done that you are forced to give number of males who agree completely the value 16, because there is a total of 36 who agree completely. Following this reason you observe that the value of the last column is forced upon you by the column-totals and similarly the value of the last row is forced upon you by the row-totals. From the original 5x2 table you can thus choose the values of 4x1 cells fairly freely (they must all be positive integers!) giving you 4 degrees of freedom!

The observed χ^2 -value is 9.175 and the expected χ^2 -value is thus 4. This means that the observed χ^2 -value is a little more than the expected value. This is not terribly critical, but we will need more information to decide what to do. But if e.g. the observed value had fallen below the expected value we would immediately have concluded that the observed difference was not significant!

To proceed we must now simulate the null hypothesis. This requires us to reconstruct the original raw data, since up till now we have only been considering the summary data. We thus introduce two lists `sex` and `position` recording the observed combinations, so the lists start with 12 females who completely agree, then follow 37 females who almost agree etc.

A	B	C	D	E	F sex	G position	H sim_position	I
1	Observed	Female	Male	Total	Chi2_obs	Female	comp. agree	almost disagree...
2	comp. agree	12	24	36	9.175	Female	comp. agree	comp. agree
3	almost agree	37	46	83	Chi2_sim	Female	comp. agree	almost disagree...
4	neither\nor	22	20	42	4.273	Female	comp. agree	almost agree
5	almost disagree...	14	6	20		Female	comp. agree	almost disagree...
6	comp. disagree	9	5	14		Female	comp. agree	neither\nor
7	Total	94	101	195		Female	comp. agree	almost agree
8						Female	comp. agree	almost agree
9	Expected					Female	comp. agree	almost disagree...
10	comp. agree	17.35	18.65	36.		Female	comp. agree	comp. agree
11	almost agree	40.01	42.99	83.		Female	comp. agree	almost agree
12	neither\nor	20.25	21.75	42.		Female	comp. agree	almost agree
13	almost disagree...	9.641	10.36	20.		Female	almost agree	comp. disagree...
14	comp. disagree	6.749	7.251	14.		Female	almost agree	almost agree
15	Total	94.	101.	195.		Female	almost agree	neither\nor
16						Female	almost agree	comp. agree
17	Simulated					Female	almost agree	neither\nor
18	comp. agree	16	20	36		Female	almost agree	almost agree
19	almost agree	35	48	83		Female	almost agree	almost disagree...
20	neither\nor	22	20	42		Female	almost agree	almost agree
21	almost disagree...	12	8	20		Female	almost agree	almost agree

You can construct these lists manually dragging down the requested values through the appropriate number of cells. But you can also compute them automatically using some nasty formulas. For your reference I quote these formulas, but don't let them disturb your sleep at night, if you don't grasp their meaning immediately☺.

```
F sex:=augment(seq{b1,'x',1,sum{b2:b6}},seq{c1,'x',1,sum{c2:c6}})
```

```
G position:=augment(freqtable@>list{a2:a6,b2:b6},freqtable@>list{a2:a6,c2:c6})
```

Once the lists of the original raw data are put in place we can scramble the position using the now familiar command:

```
H sim_position:=randsamp(position,dim(position),1)
```

Finally we can mix the original variable `sex` with the scrambled variable `sim_position` to obtain the list of simulated index carts `sim_mix` using the cell-command

```
I1 =f1&h1
```

You fill down this cell-command along the other lists of data!

This is it! You can now simulate the null hypothesis of independence by pressing **CTRL R!**

A	B	C	D	E	F sex	G position	H sim_position	I sim_mix	J
1	Observed	Female	Male	Total	Chi2_obs	Female	comp. agree	almost disagree...	Femalealmost disagree
2	comp. agree	12	24	36	9.175	Female	comp. agree	comp. agree	Femalecomp. agree
3	almost agree	37	46	83	Chi2_sim	Female	comp. agree	almost disagree...	Femalealmost disagree
4	neither\nor	22	20	42	4.273	Female	comp. agree	almost agree	Femalealmost agree
5	almost disagree...	14	6	20		Female	comp. agree	almost disagree...	Femalealmost disagree
6	comp. disagree	9	5	14		Female	comp. agree	neither\nor	Femaleneither\nor
7	Total	94	101	195		Female	comp. agree	almost agree	Femalealmost agree
8						Female	comp. agree	almost agree	Femalealmost agree
9	Expected					Female	comp. agree	almost disagree...	Femalealmost disagree
10	comp. agree	17.35	18.65	36.		Female	comp. agree	comp. agree	Femalecomp. agree
11	almost agree	40.01	42.99	83.		Female	comp. agree	almost agree	Femalealmost agree
12	neither\nor	20.25	21.75	42.		Female	comp. agree	almost agree	Femalealmost agree
13	almost disagree...	9.641	10.36	20.		Female	almost agree	comp. disagree	Femalecomp. disagree
14	comp. disagree	6.749	7.251	14.		Female	almost agree	almost agree	Femalealmost agree
15	Total	94.	101.	195.		Female	almost agree	neither\nor	Femaleneither\nor
16						Female	almost agree	comp. agree	Femalecomp. agree
17	Simulated					Female	almost agree	neither\nor	Femaleneither\nor
18	comp. agree	16	20	36		Female	almost agree	almost agree	Femalealmost agree
19	almost agree	35	48	83		Female	almost agree	almost disagree...	Femalealmost disagree
20	neither\nor	22	20	42		Female	almost agree	almost agree	Femalealmost agree
21	almost disagree...	12	8	20		Female	almost agree	almost agree	Femalealmost agree
22	comp. disagree	9	5	14		Female	almost agree	comp. agree	Femalecomp. agree
23	Total	94	101	195		Female	almost agree	almost agree	Femalealmost agree
24						Female	almost agree	neither\nor	Femaleneither\nor
25						Female	almost agree	almost agree	Femalealmost agree

B18 =countif(sim_mix,b\$1&\$a2)

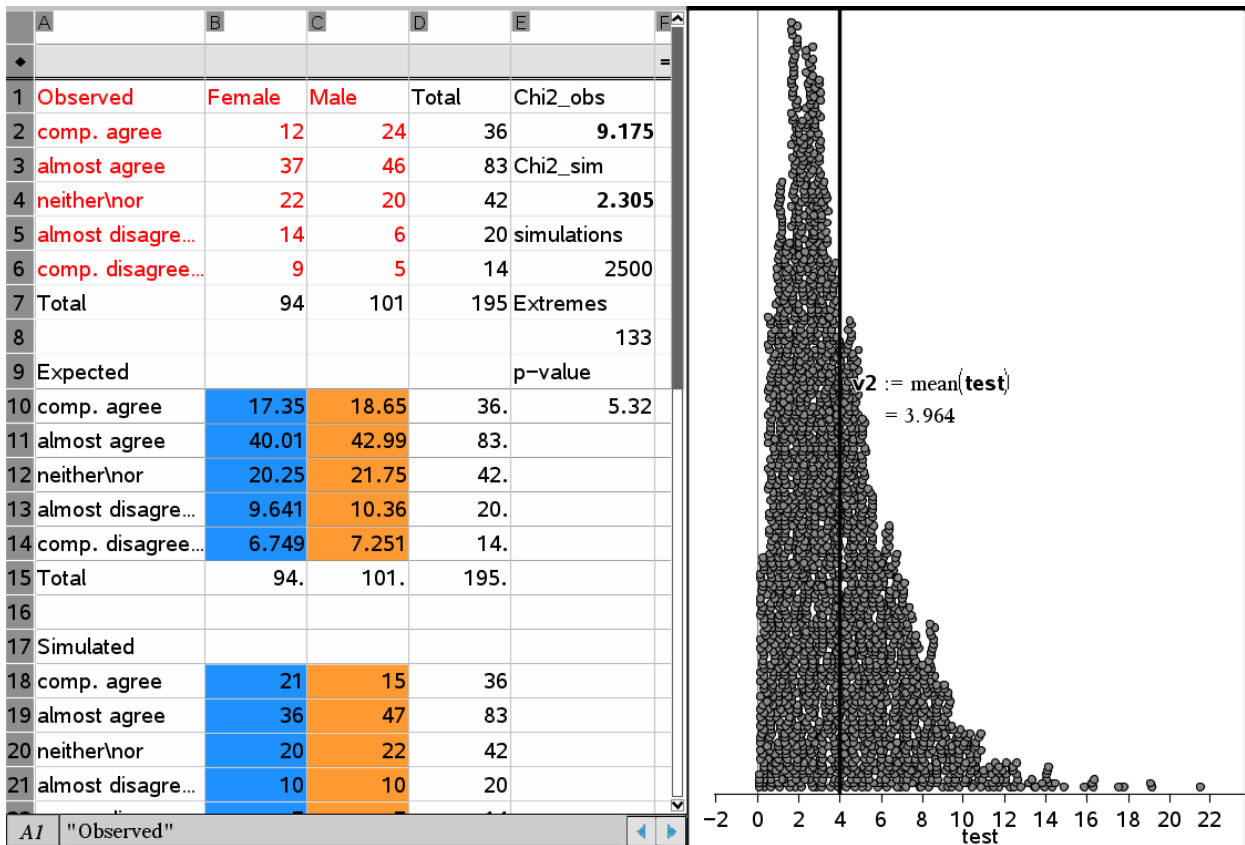
For comparison with the observed counts and the expected counts we have now added a summary table of the simulated counts. This allows you to compute the χ^2 -value of the simulation using the now familiar cell-command:

$$E4 \quad \text{chi2_sim} := \text{sum} \left(\frac{(b18:c22 - b10:c14)^2}{b10:c14} \right)$$

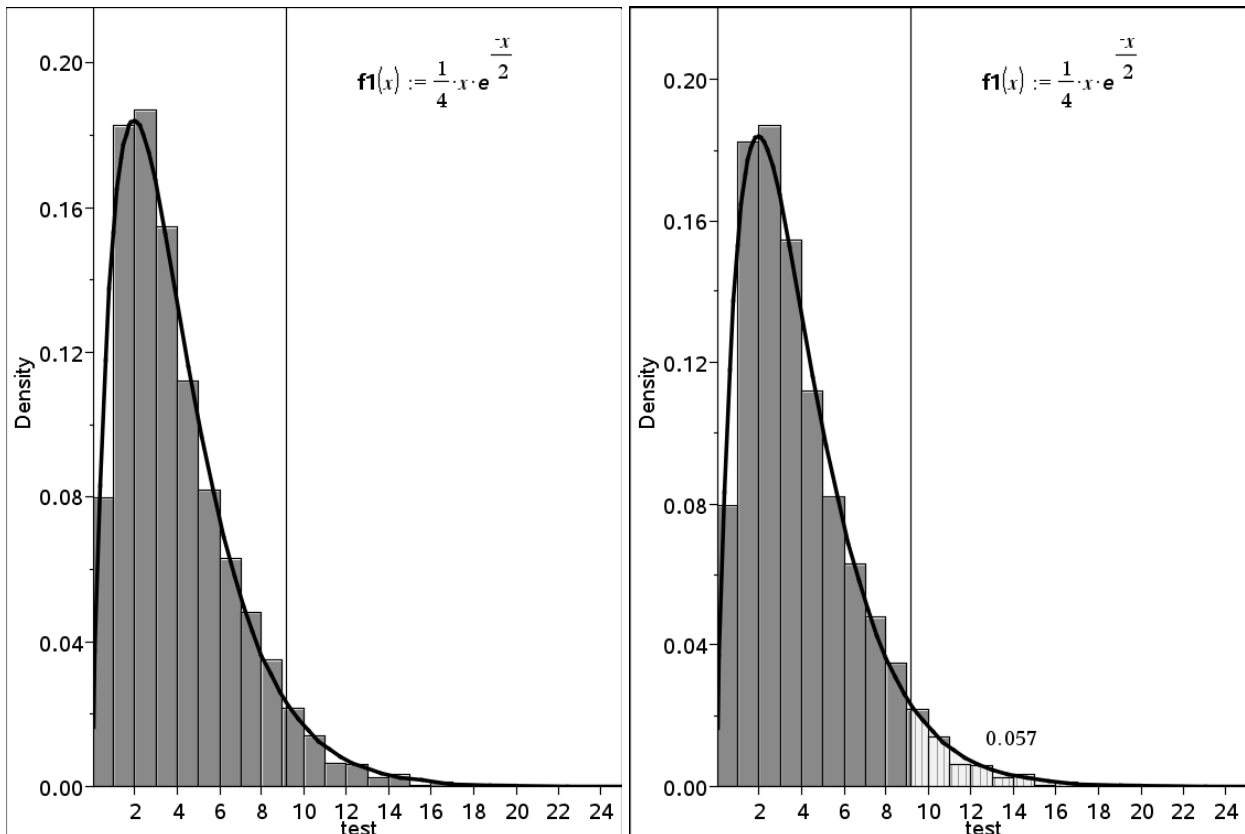
This time we will capture this using an automatic data capture. Chances are very small for having the same value twice in a row, so we can afford to ignore these rare occurrences! Setting up the data capture and capturing 2500 simulated values of the simulated χ^2 -value we see a distinct outline of the distribution emerge on the screen. We have added the mean value of the simulated test-value to the diagram. The mean value is 3.964, which is fairly close to the expected 4, that you will only get in the limit if you continues the simulation for ever 😊

It the two variables were independent (the null hypothesis) you would thus obtain an average value of 4 when you simulate a great number of times. The observed value of 9.175 is not that far away from the expected value.

But to really decide if the null hypothesis must be rejected we must now count the extremes and compute the proportion of extremes, i.e. the p-value! The p-value turns out to be 5.3%, which is greater than the standard level of significance 5%, so we cannot reject the null hypothesis. We may therefore accept that sex and position can be treated as independent variables.

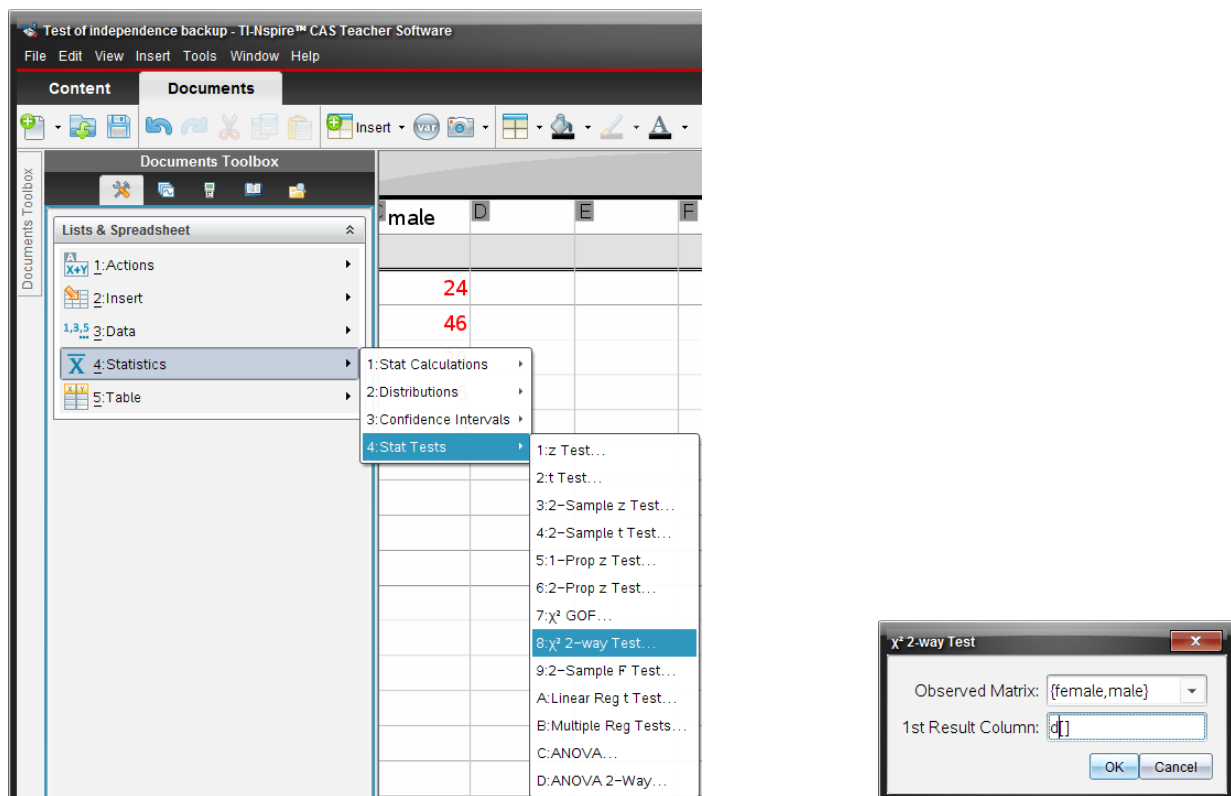


To understand the formal aspects of the χ^2 -test better we now switch to a histogram of the distribution of the test quantity. This histogram reflects that theoretical χ^2 -distribution with 4 degrees of freedom, labelled $\text{chi2Pdf}(x,4)$ in TI-Nspire™ CAS. Despite its exotic name it is just the familiar function $y = \frac{1}{4} \cdot x \cdot e^{-x/2}$.



When we want to perform the χ^2 -test of independence we therefore do not need to set up the experimental machinery of simulating the null hypothesis. We can simply replace the experimental distribution with the theoretical built in distribution and compute the p-value directly as the area cut off by the observed χ^2 -value 9.175. We then obtain the theoretical p-value 0.057, i.e. 5.7% leading to the same conclusion.

Finally it should be pointed out that the χ^2 -test is of course built in as a standard test in TI-Nspire™ CAS. To perform the test all you need is the original summary table, but you need it entered as named lists 'above the bar', so that you can refer to the lists in the dialogue box, which unfortunately do not accept cell ranges:



	A position	B female	C male	D	E
◆					= χ^2 2way({female,male}): Cop
1	comp. agree	12	24	Title	χ^2 2-way Test
2	almost agree	37	46	χ^2	9.175
3	neither\nor	22	20	PVal	0.0569
4	almost disagree	14	6	df	4.
5	comp. disagree	9	5	ExpMatr...	[[17.353846153846,40.0102..
6				CompMa...	[[1.6517184942717,0.22648..

This makes it possible to perform the test very easy and obtain the theoretical p-value 0.0569 without any effort. Notice however that to use this test you must have a deep understanding of the meaning of the test. And a lot experiences with teaching inference statistics shows that this understanding most easily comes from getting acquainted with simulations of null hypothesis. The experimental methods thus functions as ladder, that allows you to climb up to the next level, where you can perform automatic testing with great confidence 😊

4. Conclusions

When teaching inferential statistics understanding the null hypothesis is a sine qua non. And one way of obtaining this understanding is through the simulation of null hypothesis. Another sine qua non is the understanding of the distribution of the test statistics which is approximated by the theoretical point distribution function. Again this understanding is facilitated by the experimental approach. The experimental approach however requires some training before you handle it adequately, so there is still a lot of work to do if you want to integrate it in the teaching of statistics. In Denmark the experimental approach is optional, but is accepted as a viable alternative to the more traditional theoretical approach. It is my belief that students will benefit from the experimental approach, so that is worth the efforts required.

Another issue is the choice of χ^2 -tests. What about other tests like t-tests etc? Is it really worth focusing so much upon the χ^2 -test? First it delimits a precise area within the inferential statistics that you can test at the written exam. Second once you understand the machinery behind the χ^2 -test experience shows that it is relatively easy to switch to other tests, which can now be performed automatically with confidence, since the underlying concepts are the same: You need to understand the null hypothesis and you need to obtain a p-value. It is only the first test you have to work really hard for, the others come for free 😊

But as the adjustment of the Danish Curriculum has just been performed recently no students have yet completed the new curriculum. Only time will tell whether the transition will be smooth or we will run into unforeseen obstacles.

References

David Salsburg: The Lady Tasting Tea – How Statistics revolutionized Science in the Twentieth century, 2001.

Ronald Fisher: The design of Experiments, 1935.