**A simple explanation to understand Chi-Square Test**

[[](https://medium.com/@wenyiyan?source=post_page-----1814fa261499----------------------)](https://medium.com/@wenyiyan?source=post_page-----1814fa261499----------------------)

[Wenyi YAN](https://medium.com/@wenyiyan?source=post_page-----1814fa261499----------------------)

[Jan 22](https://medium.com/wenyi-yan/a-simple-explanation-to-understand-chi-square-test-1814fa261499?source=post_page-----1814fa261499----------------------) · 6 min read

Let’s use daily life examples to explain and understand what is a Chi-Squared test.

Chi-squared test, a statistical method, is used by machine learning methods to check the correlation between two categorical variables.

Chinese people translate Chi-Squared test into “card-squared” test, which makes most people think, this methodology is related to a card, and come up with a question.

“Do I need to buy a card to do this test? “

Of course, not.

I often hear this conversation between product managers and analysts.

Analyst: “I did a Chi-Squared test for this variable. The result shows that the variable is not significant, so I didn’t put it into the final model.”

At this time, many project managers may not understand what a Chi-Squared test is.

Some “good students” will ask the analyst directly “Hi, Can you tell me what is the Chi-Squared test? ”

Others might sneakily check out Wikipedia to find out what is a Chi-Squared test >\_<.

Either way, most product managers will take the analyst’s advice with no further questions or discussions.

This actually happened to me on a daily basis. In my view, product managers main job is planning and execution throughout the product life-cycle, based on input and recommendations from data scientists/analysts, not working on the problems themselves.

The objectives of this article are:

1. To tell the product manager: don’t be afraid when you hear this word!
2. To use simple examples to make the statistical concept more friendly for beginners.
3. To teach analysts to say ‘layman’ words in conversations with cross-functional teams.

What is the Chi-Square test:

Chi-square test evaluates if two categorical variables are related in any way.

We can use it in the following scenarios:

1. Test if the gender places a significant difference role on the online grocery shopping decision.
2. Test if the city tier places a significant difference in the segment of car purchase decision, etc.

If there is a significant difference, we will consider putting these variables into the model or analysis.

Coin

Let’s start with one of the simplest examples.

1. Determine whether the coin is fair or unfair based on the # of heads and # of tails when you toss the coin.



Okay, let me ask this question in another way. Let’s say, if I give you a normal coin. Remember! It is a normal coil with one head and one tail. How many # of heads and # of tails will you get if you toss it 50 times?

Based on logic, the best case is 25 heads and 25 tails.

But I don't believe it will get the perfect scenario of 25 heads vs 25 tails.

Then, you will start to think, 28 heads and 22 tails are okay, and 23 heads and 27 tails are fine as well.

But you won’t believe you will get 10 heads and 40 tails from a normal coin. If that happens, you must think you should buy the lottery immediately!

The above thinking process is for you to take the result (the coin is normal), and guess the different cases that would occur.

The Chi-Squared test is just to reverse the process of the above thinking process.

To make a conclusion by the observation (To determine the # of heads, # of tails)

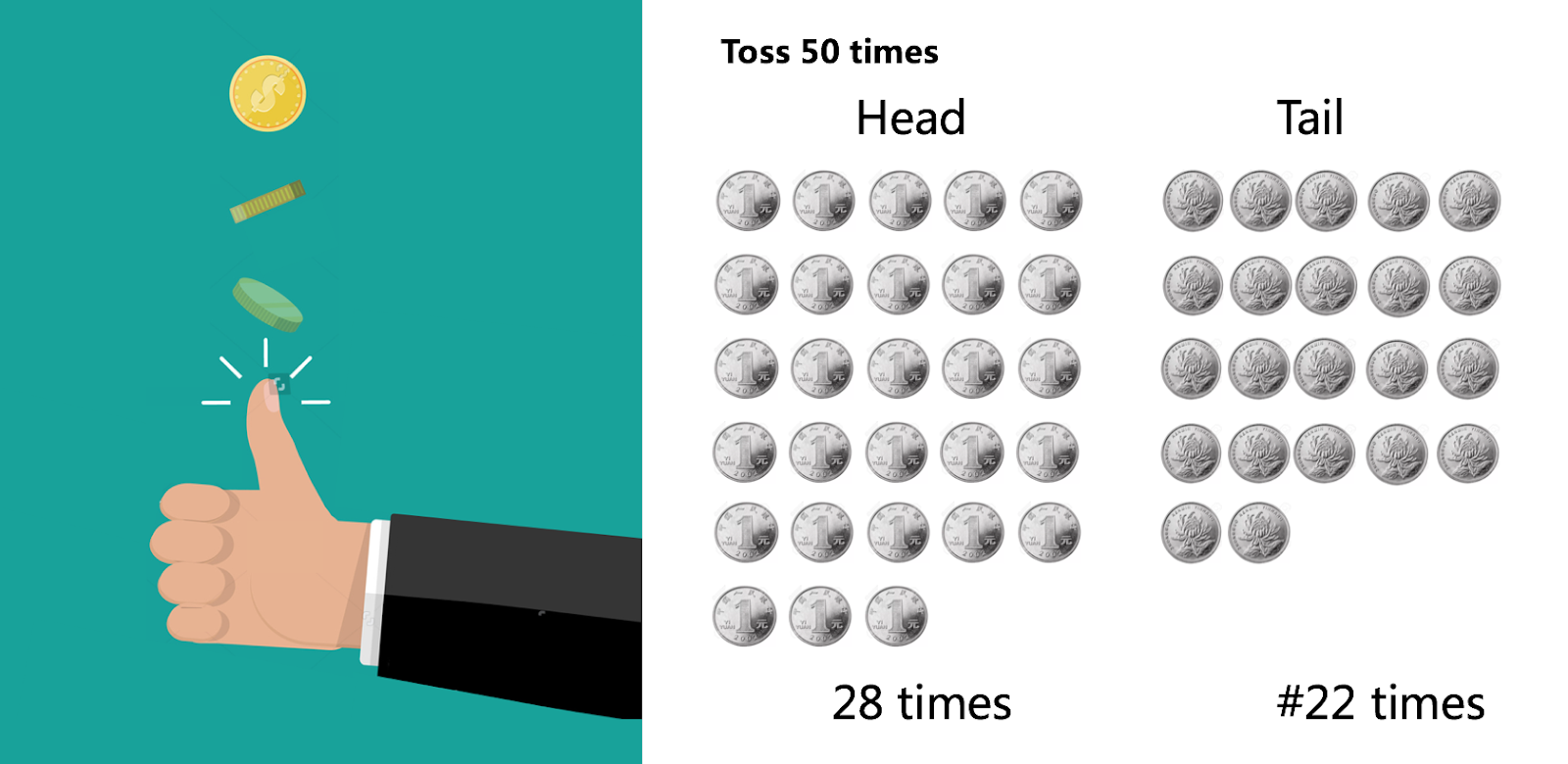
Let’s go back to the example.

If I don’t know the coin’s outcome. I would like to check the # of heads and # of tails.

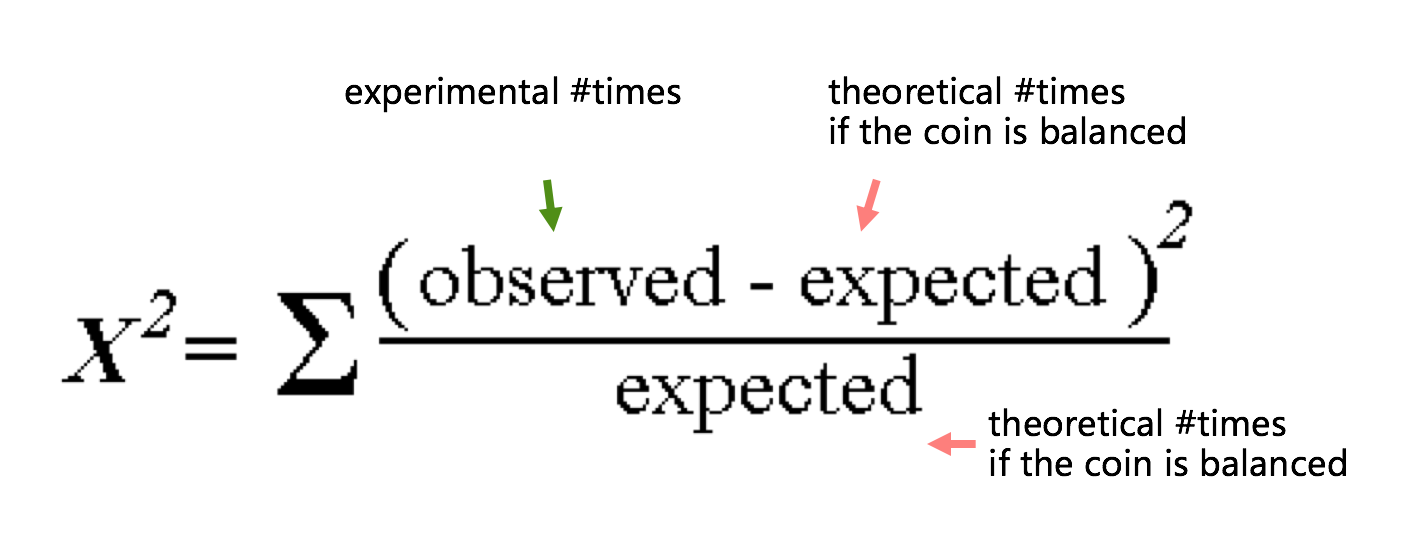
I started my experiment and tossed the coin 50 times.

Then I got 28 heads and 22 tails.

How should I use the Chi-Squared test to determine the outcome of the coin?



Here is the formula for the Chi-Squared test:



This formula can help us determine the score of the Chi-Squared test.

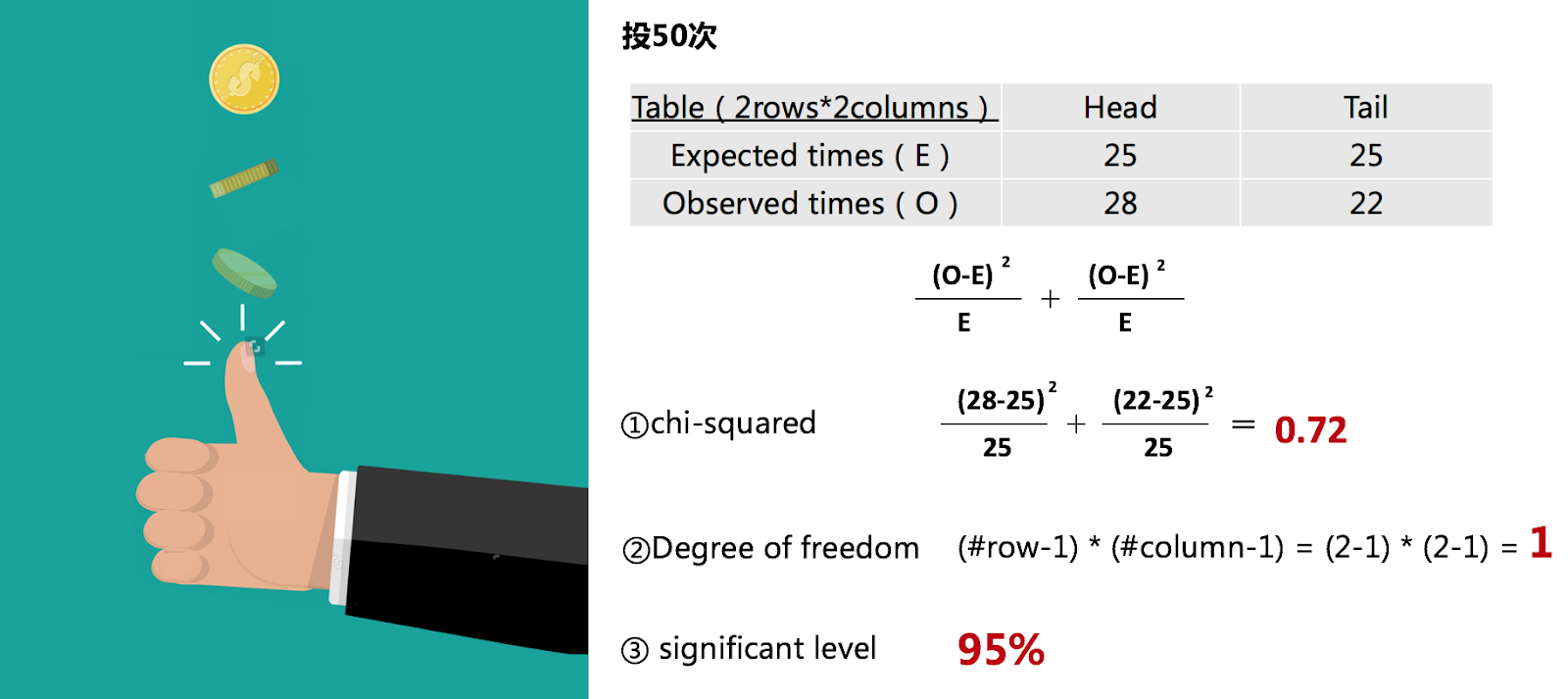
The information I need:

1. The value of x²

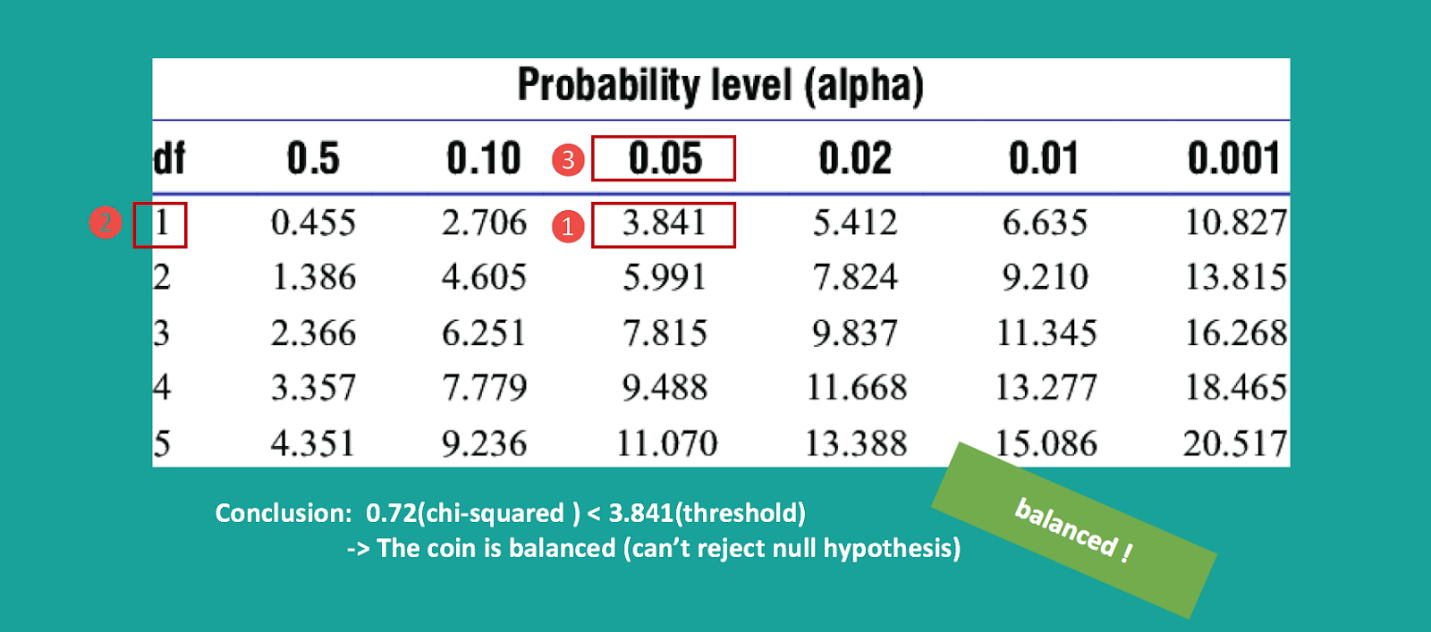
2. The Degrees of Freedom (if you don’t know what is DoF, please check my DoF article in my blog)

3. Confidence Level (usually 90% or 95%)

We put the 3 values into the formula and we get,



Once we get these three values then we could look up Chi-Squared table, because 0.72 is less than 3.84, we can not reject null hypothesis, so we can make the conclusion that the coin is fair.



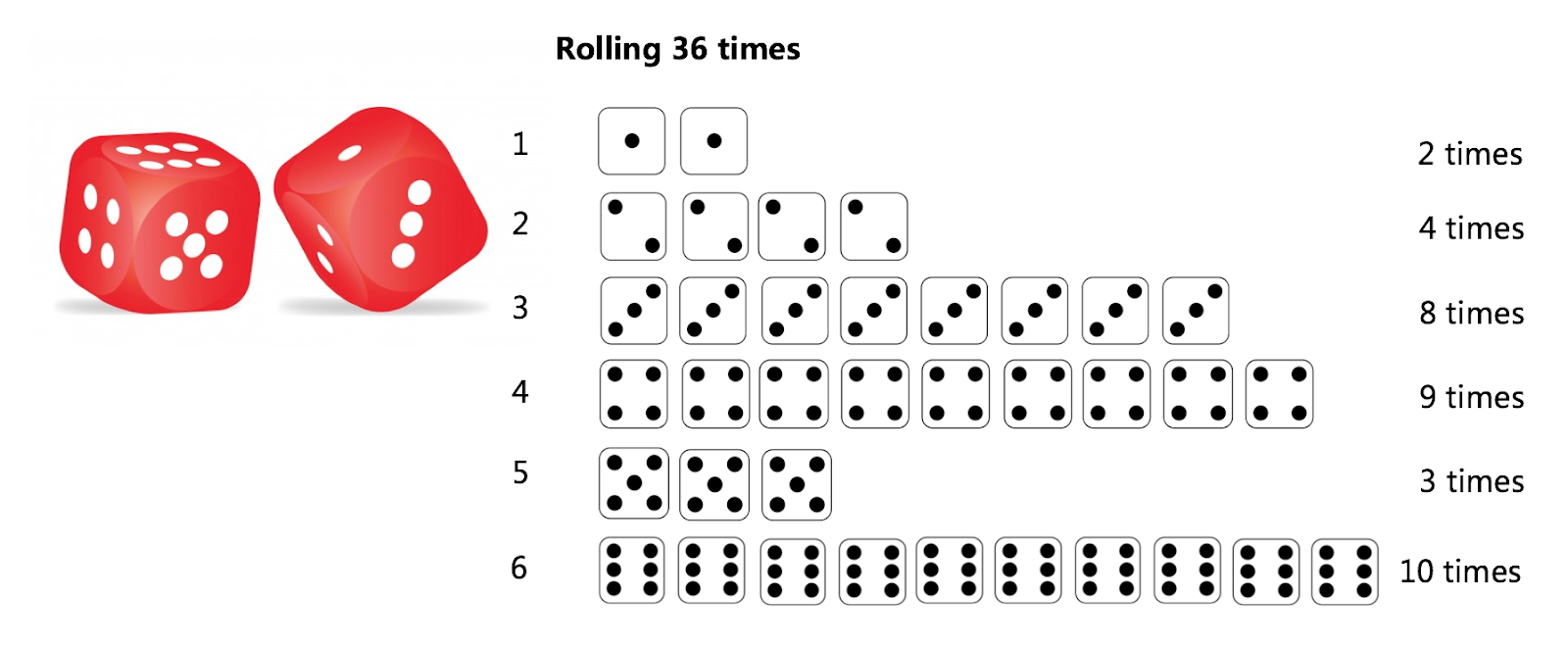
(For the hypothesis test, please check my hypothesis blog)

If you are confused after checking the lookup table, you should ask yourself one more question: what if the coin is fair, the Chi-Squared value should be smaller or bigger?

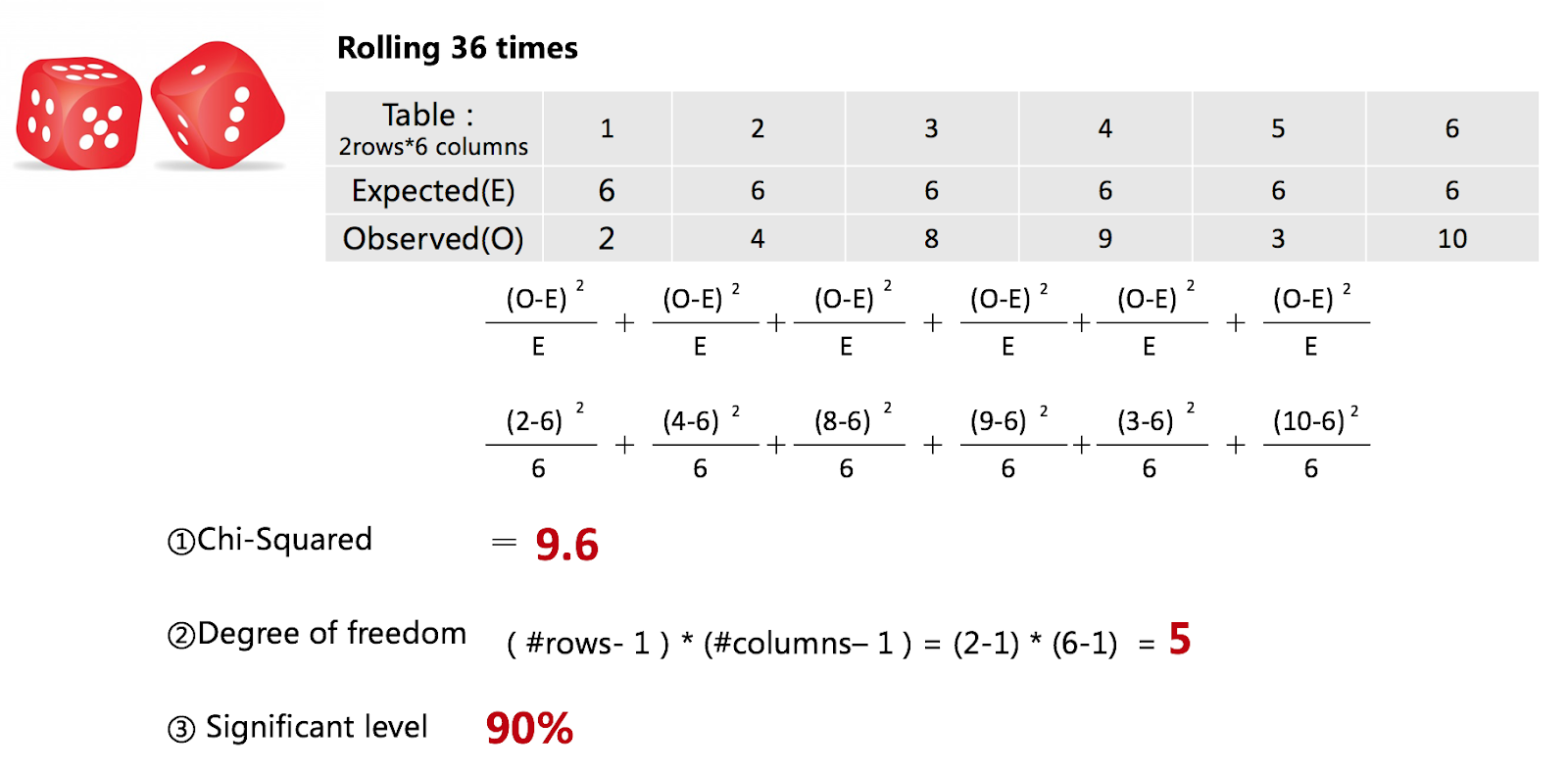
Rolling a Dice

Next, let’s look at a more difficult example: rolling a dice.

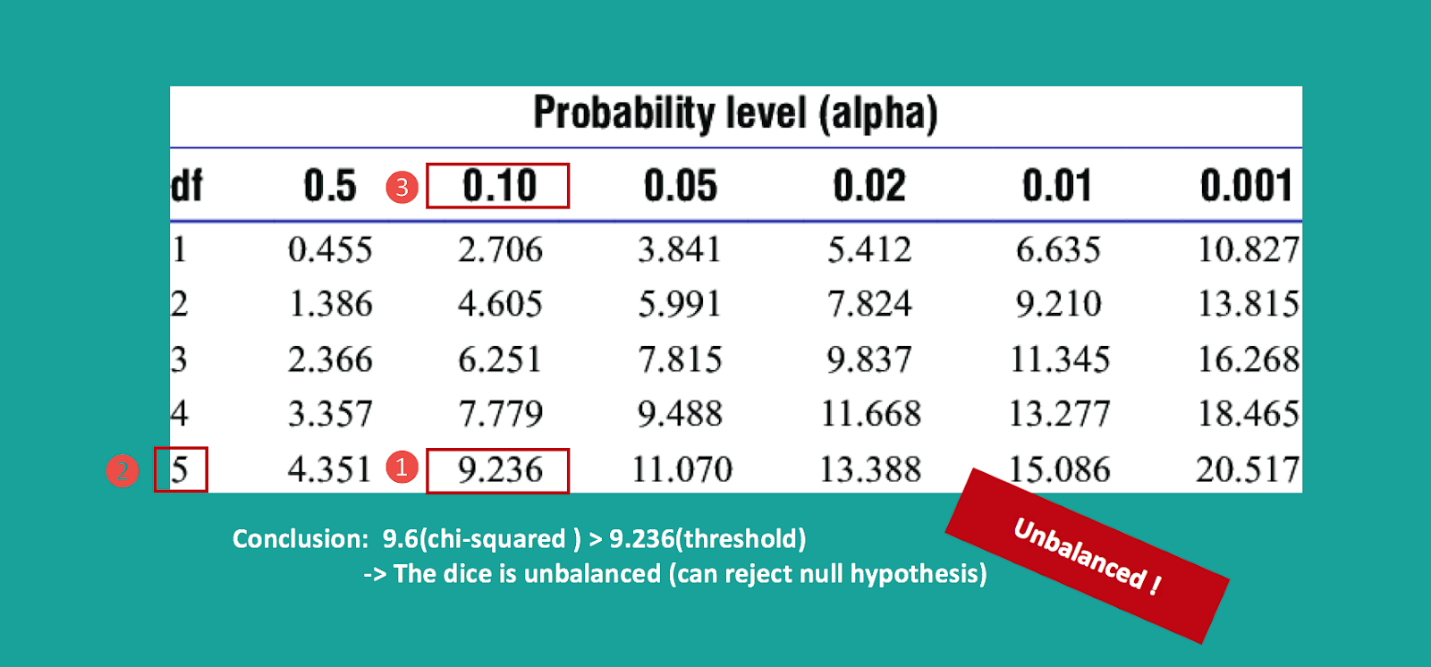
There is a dice, I don’t know if it is a fair dice so I plan to roll it 36 times.



According to the way of tossing the coin, I have to draw a table and then calculate 3 values.



With these three values, we can go ahead to check the lookup table, so we can not accept that this is a fair dice, and we have to say this is an unfair dice.



So far, you understand that the Chi-Squared test is not that esoteric!

When you discuss a report with data analyst or data scientist next time, they tell you: hey, this variable is not significant, I removed from this analysis.

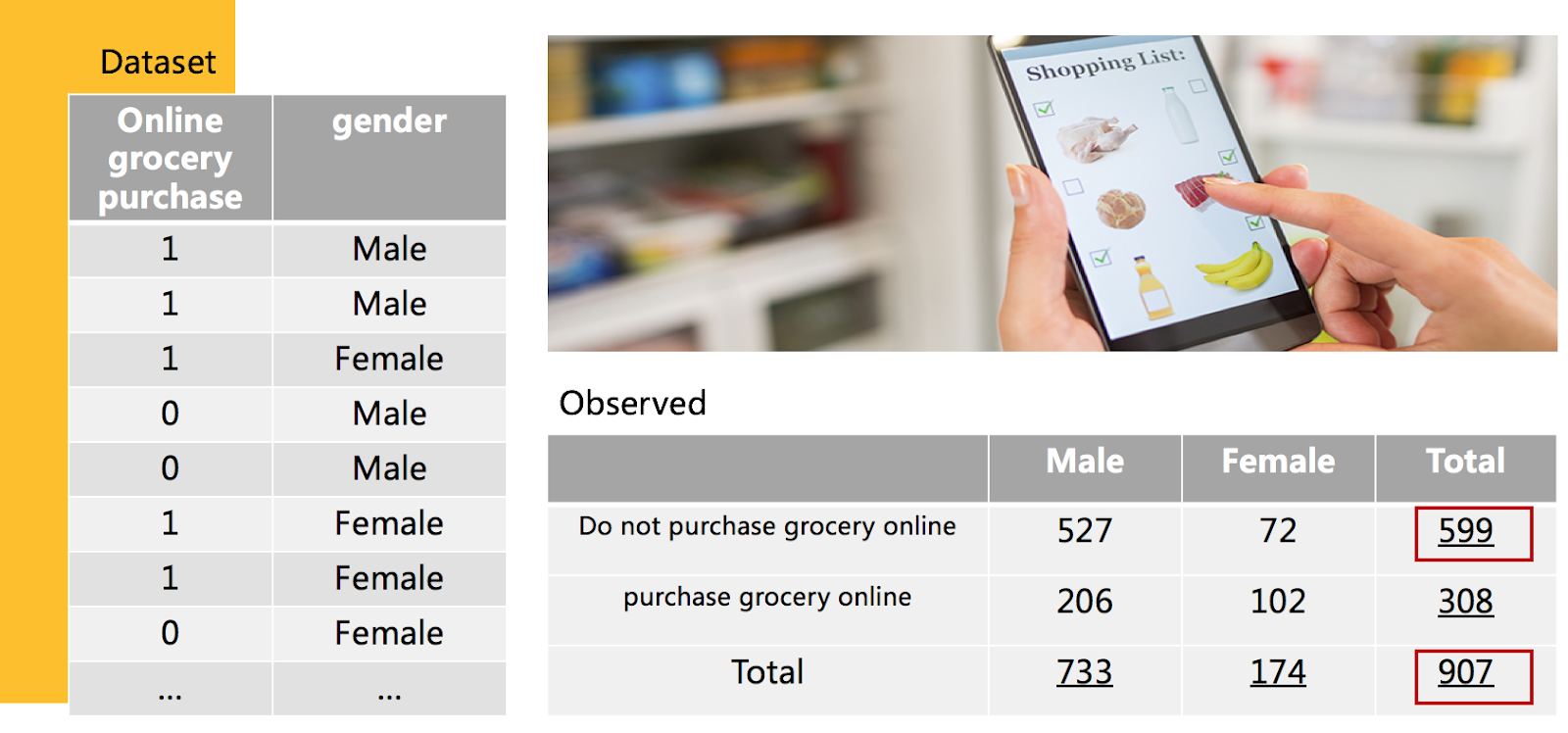
Then, you can ask him with confidence, what is the Chi-Squared value? What’s your confidence level did you choose?

E-Commerce Gender vs. Purchase Online

Finally, let me talk about an application in a real use case:

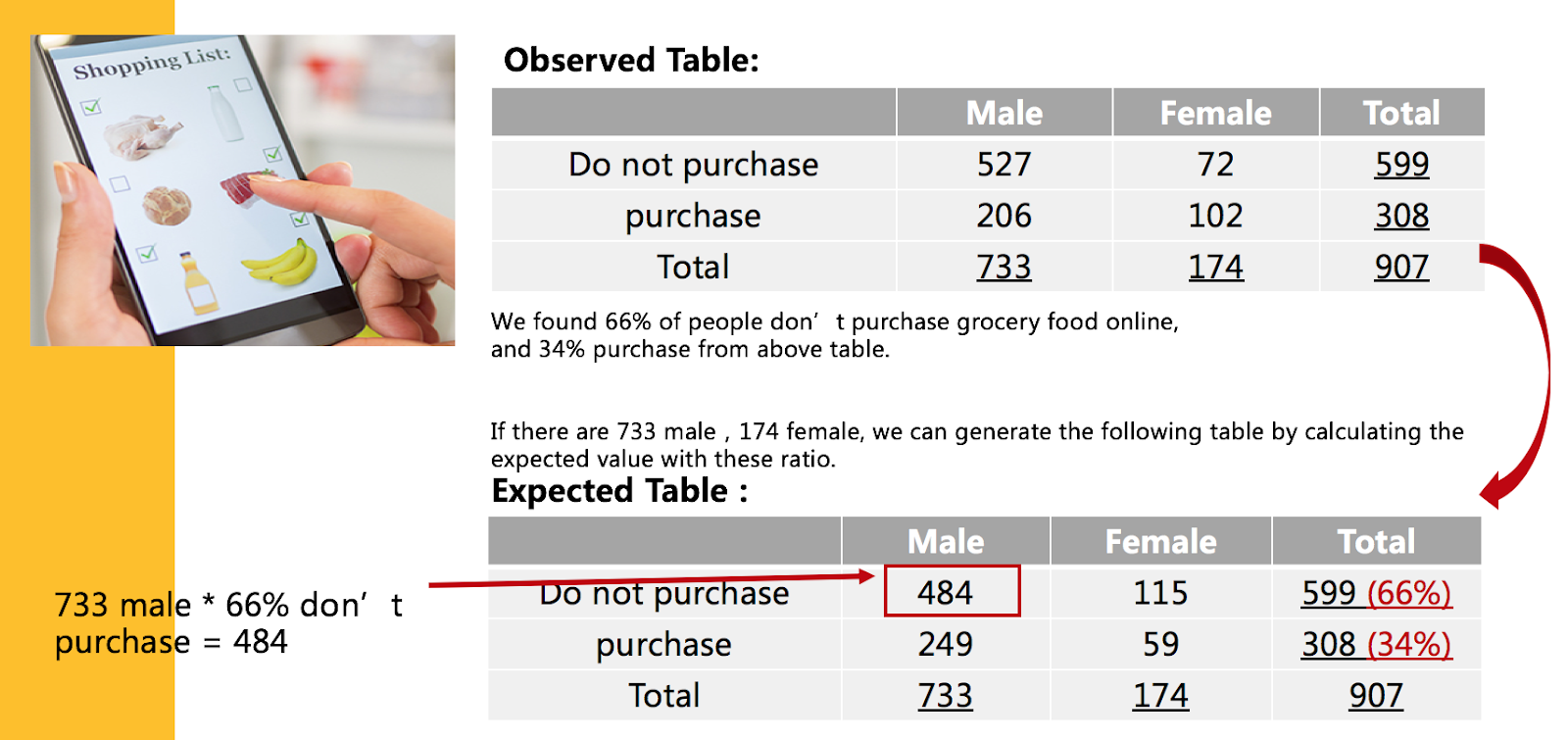
We have to test whether gender plays a significant role in the online purchase decision. In real life, women usually go to the grocery store to buy grocery. So is online the same?

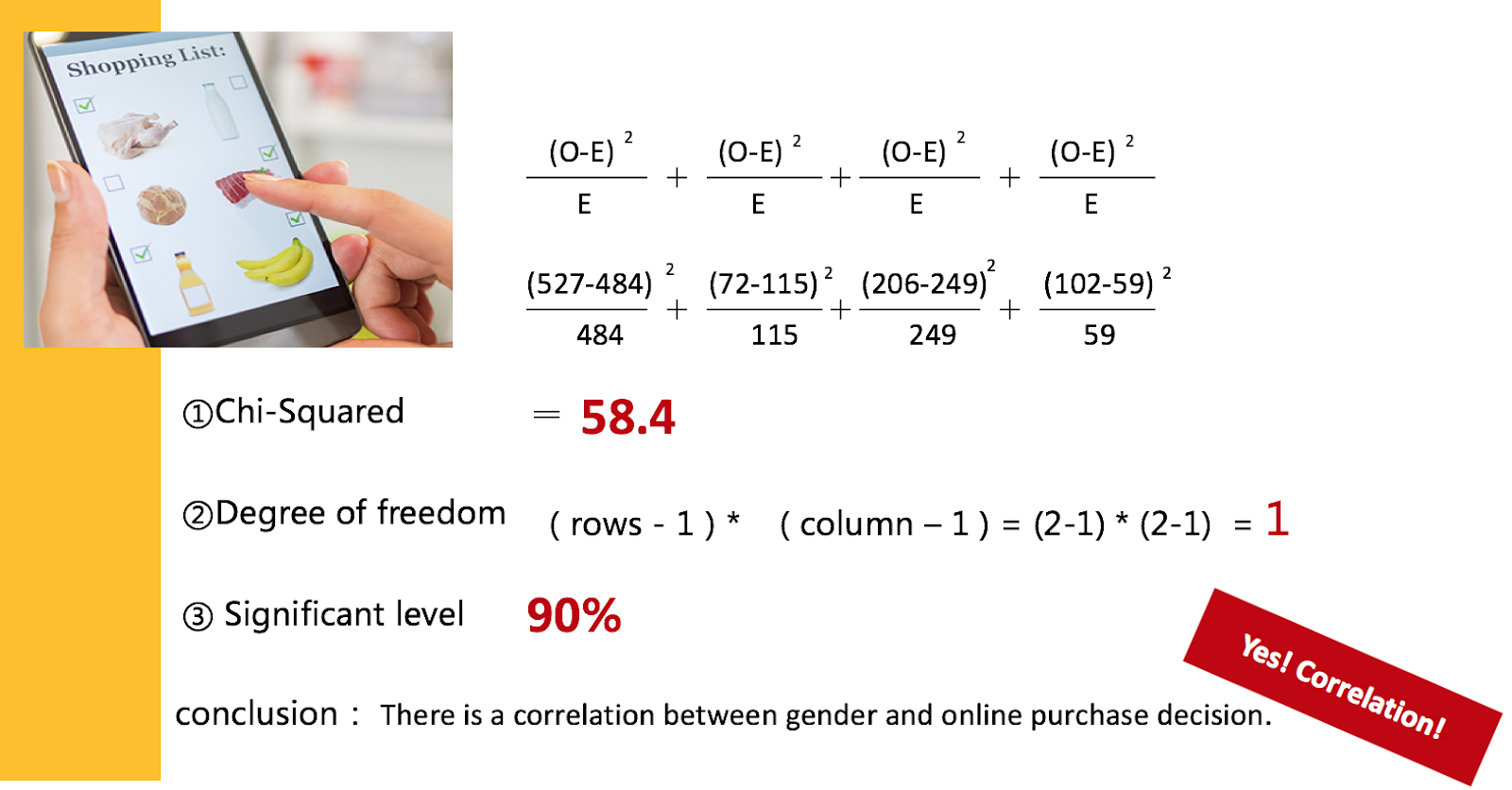
Let’s repeat the process as above,



From the table above, we can calculate 66% (599/907 ~ 66%) of people do not purchase food online. 34% of people do purchase online.

In this case, if we collect an online decision of 733 men and 174 women, how to get the Chi-Squared value?





Based on the expected and observed values, we can calculate the Chi-Squared value, along with the degree of freedom, and the confidence level which we defined.

After we check the lookup table, we can get the conclusion that the gender is significantly related to the online purchase decision.

So, if we see a woman is browsing our website next time, let try to retarget them by launching specific campaigns.

**After reading these examples, I believe that you will think Chi-Squared test is not that complicated as we thought.**

Please check : <https://www.jianshu.com/p/807b2c2bfd9b> for Chinese version ：）