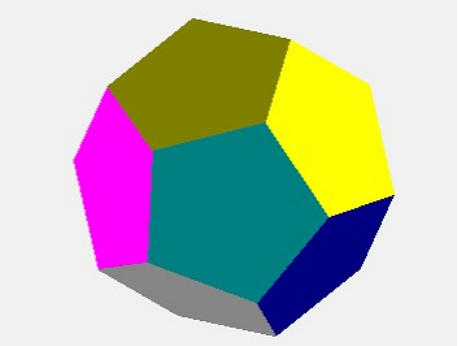
Probability Theory

An introduction and beyond

Chapter 7

Fitting data to theory

This is an article from my home page: www.olewitthansen.dk



Ole Witt-Hansen (2017) 1997

Contents

1. The theoretical foundation of regression **3**
1.1 The method of maximum likelihood 3

2. Linear regression **4**

4. Correlation coefficients **5**

5. Chi-square test **6**

6. The chi-square density function and distribution function **7**

7. Pearson's formula for the χ^2 test **9**

8. Explaining Pearson's formula. **11**

1. The theoretical foundation of regression

The method of least squares supplies the standard method if you want to fit experimental data to some theoretical model. Say, that we have a series of measurements $(x_1, x_2, x_3, ..., x_n)$, and we want to compare it to some function $y = f(t, \alpha_k)$, where $y_i = f(t_i, \alpha_k)$, and the parameters α_k are to be determined to give the best fit. The method is to minimize the sum:

$$\chi^2 = \sum_{i=1}^{n} (x_i - y_i(\alpha_k))^2$$

With respect to the parameters α_k.

The method of least squares is a special case of the more general: *Method of maximum likelihood,* which already had been applied by Gauss, and later has been extensively described by R.A. Fischer.

1.1 The method of maximum likelihood

Let *A* and *B* be two equal likely theories, and let *E* be an experiment, which is conducted to decide which theory is the most reliable. According to Bayes theorem:

$$P(A|E)P(E) = P(E|A)P(A) \quad \text{and} \quad P(B|E)P(E) = P(E|B)P(B)$$

According to the premise $P(A) = P(B)$, we then find by division of the two equations.

$$\frac{P(A|E)}{P(B|E)} = \frac{P(E|A)}{P(E|B)}$$

This equation merely expresses that the most probable theory, given the experiment, is the theory that gives the most probable outcome of the experiment, which is hardly surprising.

Let us assume that we have *n* independent measurements $x_1, x_2, x_3, ..., x_n$, which have a probability distribution $P(X, \alpha)$ that also depends on some parameter α, which we want to determine.

Since the observations are considered to be independent, then the probability of the outcome of the experiment is the product of the probabilities of each of the independent measurements.

The result is the so called Likelihood function

$$L(\alpha) = P(x_1, \alpha)P(x_2, \alpha)P(x_3, \alpha) \cdots P(x_n, \alpha)$$

The ”most likely theory” is, according to Bayes theorem, the values of α, which maximizes $L(\alpha)$, which is the cause of the name ”Method of Maximum Likelihood”.

For practical reasons one takes logarithm to $L(\alpha)$, and determines the maximum of $-\ln L(\alpha)$ (since $\ln(x) < 0$ for $0 < x < 1$)

Assuming specifically that $x_1, x_2, x_3, .., x_n$ are normally distributed with the same mean μ(α), and standard deviation σ_i, such that

$$P(x_i, \alpha) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{\frac{(x_i - \mu(\alpha))^2}{2\sigma_i^2}} \quad \Rightarrow \quad \ln P(x_i, \alpha) = -\ln\sqrt{2\pi}\sigma_i - \frac{(x_i - \mu(\alpha))^2}{2\sigma_i^2}$$

If we discard all factors, which are unimportant for determination of the max of $-\ln L(\alpha)$, we have:

$$L(\alpha) = \frac{(x_1 - \mu(\alpha))^2}{2\sigma_1^{\,2}} + \frac{(x_2 - \mu(\alpha))^2}{2\sigma_2^{\,2}} + \frac{(x_3 - \mu(\alpha))^2}{2\sigma_3^{\,2}} + \ldots + \frac{(x_n - \mu(\alpha))^2}{2\sigma_n^{\,2}}$$

If all measurements have the same theoretical standard deviation (and only then), we find, (when we multiply the expression with $2\sigma^2$), the familiar expression for χ^2 (Chi-square)

$$\chi^2 = (x_1 - \mu(\alpha))^2 + (x_2 - \mu(\alpha))^2 + (x_3 - \mu(\alpha))^2 + \ldots (x_n - \mu(\alpha))^2$$

or

$$\chi^2 = \sum_{i=1}^{n} (x_i - \mu(\alpha))^2$$

Finding the minimum value for this expression, means the least over all deviation of x_i from $\mu(\alpha)$, and therefore the best fit.

This is done by solving the equation $\frac{\partial \chi^2}{\partial \alpha} = 0$ (Since χ^2 can have no max)

This result is trivially generalized to more than one parameter, and more than one variable.

2. Linear regression

We shall now specialize to the case, where we have a stochastic variable *Y*, which according to some theory depends linearly on *x*. Specifically: $y_i = a \cdot x_i + b$.
The theoretical mean of *y* should be $E(y) = aE(x) + E(b)$ or $\bar{y} = a\bar{x} + b$. We shall further assume that the y_i are normally distributed, and that they have the same variance. This is not always the case, but it is the presupposition of using the method of least squares.

As described above: “The principle of maximum likelihood” leads us toward finding the minimum of the sum of squares as a function $f(a,b)$, of the parameters *a* and *b*.

$$f(a,b) = (y_1 - (ax_1 + b))^2 + (y_2 - (ax_2 + b))^2 + (y_3 - (ax_3 + b))^2 + \ldots + (y_n - (ax_n + b))^2$$

Writing it more compactly:

$$f(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

By evaluating $\frac{\partial f}{\partial a}$ and $\frac{\partial f}{\partial b}$ and put them equal to zero, we obtain two equations, where the summation from $i = 1$ to n is implied.

$$\frac{\partial f}{\partial a} = \sum -2x_i(y_i - ax_i - b) = 0 \qquad \text{and} \qquad \frac{\partial f}{\partial b} = \sum -2(y_i - ax_i - b) = 0$$

Doing a bit of algebra on both equations we find:

$$a\sum x_i^{\,2} + b\sum x_i = \sum x_i y_i \qquad \text{and} \qquad a\sum x_i + nb = \sum y_i$$

From the last equation follows:

$$b = \frac{1}{n}\sum y_i - a\frac{1}{n}\sum x_i = \bar{y} - a\bar{x} \qquad \text{where} \qquad \bar{y} = \frac{1}{n}\sum y_i \quad \text{and} \quad \bar{x} = \frac{1}{n}\sum x_i$$

When inserting this expression for b in the first equation we can isolate a.

$$a\sum x_i^{\,2} + (\bar{y} - a\bar{x})\sum x_i = \sum x_i y_i \;\Leftrightarrow\; a\left(\sum x_i^{\,2} - n\bar{x}^2\right) = \sum x_i y_i - n\bar{x}\bar{y}$$

To get:

$$a = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^{\,2} - n\bar{x}^2} \qquad \text{and} \qquad b = \bar{y} - a\bar{x} \quad (\text{ since } \quad b = \frac{1}{n}\sum y_i - a\frac{1}{n}\sum x_i \quad)$$

4. Correlation coefficients

When you have made a χ^2 fit, you are usually interested in how good the fit is.
This can be done using the χ^2 probability distribution function, but then we are engaged in rather heavy statistical mathematics. Or one may evaluate the so called correlation coefficient.
Usually one fits data to theory. If we have a series of measurements ($x_1, x_2, x_3, \ldots, x_n$) and the corresponding theory gives the result ($y_1, y_2, y_3, \ldots y_n$), one usually makes a χ^2 fit, but the question remains how much the two data sets are correlated. In this context, one usually uses, (instead of x_i and y_i), the variables $x_i - \bar{x}$ and $y_i - \bar{y}$, since they both have the mean 0.
So we want to have a quantity that measures how good the correlation between the two datasets is. The *correlation coefficient* is such a quantity.

First we define some statistical quantities, of which we shall not try to substantiate their statistical significance.

$$s_x^{\,2} = \frac{1}{n-1}\sum (x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum x_i^{\,2} - n\bar{x}^2\right)$$
$$s_y^{\,2} = \frac{1}{n-1}\sum (y_i - \bar{y})^2 = \frac{1}{n-1}\left(\sum y_i^{\,2} - n\bar{y}^2\right)$$
$$s_{xy} = \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum x_i y_i - n\bar{x}\bar{y}\right)$$

The correlation coefficient r_{xy} is then defined by:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

The correlation coefficient is, due to the Cauchy Schwartz inequality, situated in the interval [-1,1].
To get a hint of the significance of the correlation coefficient we may perceive the sets $(x_i - \bar{x})$ and $(y_i - \bar{y})$, as the coordinates to two vectors $\bar{a}$ and $\bar{b}$ in a n-dimensional space. In this interpretation it is seen that the correlation coefficient is equal to cosine between the two vectors.

$$r_{xy} = \cos v = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|}$$

If $|r_{xy}| = 1$, then $v = 0$ the two vectors are parallel and the one set $(y_i - \bar{y})$ is proportional to the other $(x_i - \bar{x})$. So y is a linear function of x, and they are 100% correlated.
On the other hand if $r_{xy} = 0$, the two vectors $(y_i - \bar{y})$ and $(x_i - \bar{x})$, are orthogonal, which we would interpret as there are no correlation between them at all.

We shall use the ”angle between the two vectors” picture to try to grasp what a certain value of the correlation coefficient implicates.
Students (and even some teachers), who have very little understanding of what a certain correlation coefficient implies, often consider that a correlation coefficient of 0.90 indicates a high correlation. However cos $v = 0.90$ means that the two datasets are separated by an angle of 25.84^0.Two vectors having this angle can hardly be considered to be parallel. Even a correlation coefficient of 0.95 gives an angle of 18.19^0, which is neither a persuasive argument for parallelism. Actually, you should have a correlation coefficient of 0.99 corresponding to an angle of $8{,}10^0$, before one should make a conclusive decision of a strong correlation.

5. Chi-square test

As mentioned above the chi-square is the quadratic sum:

$$\chi^2 = \sum_{i=1}^{n} (x_i - y_i)^2$$

Where x_i and y_i most often are the experimental and theoretical data.

The expression for Chi-square is derived from the normal distribution, where all variables have the same mean and same standard deviation. This is however often ignored, because both are usually unknown, and most variables in real life, that are subject to fluctuations are assumed (correctly) to be normally distributed.
χ^2 has a probability density function, which is not trivial. We shall do so in the next section, but here we merely state the result. One point, however, N is not (necessarily) the number of terms in the sum of squares, but the number of terms minus the number of linear relation between the variables. N is called the *degrees of freedom*.

The probability of having a result of χ^2 in the interval $[\chi^2, \chi^2 + d\chi^2]$ is given by:

$$f(\chi^2) = \frac{1}{2^{N/2}\Gamma(N/2)} e^{-\chi^2/2} (\chi^2)^{(N/2)-1} d\chi^2$$

From this follows the χ^2 probability distribution function

$$F(t) = P(\chi^2 > t) = \frac{1}{2^{N/2}\Gamma(N/2)} \int_t^{\infty} e^{-\chi^2/2} (\chi^2)^{(N/2)-1} d\chi^2$$

If one makes a Chi-square test, at a level of significance 5%, meaning that χ^2 should be in an interval [0, *t*] where statistically 95% of the outcomes should be. That is $P(\chi^2 > t) > 0.95$ or $P(\chi^2 < t) < 0.05$. There should be less than 5% probability to get a value of χ^2 lower than *t*. We should therefore determine *t*, such that $F(t) > 0.95$.
Then we should then compare χ^2 to *t*, and if $\chi^2 < t$, I has passed the test otherwise not

6. The chi-square density function and distribution function

We shall elaborate somewhat further on the chi-square density function, the expression of which we listed above. As formerly explained the chi-square density function:

$$f(\chi^2) = \frac{1}{2^{N/2}\Gamma(N/2)} e^{-\chi^2/2} (\chi^2)^{(N/2)-1} d\chi^2$$

Is an extension of the normal distribution, when having an independent series of events, having the mean μ and standard deviation σ.

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Thus having *n* normally distributed measurements $x_1, x_2, x_3, \ldots x_n$, where the standard deviations $\sigma_1, \sigma_2, \sigma_3, \ldots \sigma_n$ are known, and μ_i are the theoretical mean values, then the probability of getting a result in the infinitesimal volume $dx_1 dx_2 dx_3 \ldots dx_n$ equal to.

$$P(x_1, x_2, x_3, \ldots, x_n) dx_1 dx_2 dx_3 \ldots dx_n = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}} dx_i$$

Introducing the normalized variables q_i by $q_i = \frac{x_i - \mu_i}{\sigma_i}$ and the density function *Q* by

$$Q(q_1, q_2, q_3, \ldots, q_n) dq_1 dq_2 dq_3 \ldots dq_n = P(x_1, x_2, x_3, \ldots, x_n) dx_1 dx_2 dx_3 \ldots dx_n$$

We find:

$$Q(q) = Q(q_1, q_2, q_3, \ldots, q_n) = \left(\frac{1}{2\pi}\right)^{N/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{N} q_i^2\right)$$

We should notice that *Q* depends only on the variables q_i through the sum: $\chi^2 = \sum_{i=1}^{N} q_i^2$.

The Chi-square applies to determine whether the result of *N* measurements lie within the statistical uncertainty. The expression for *Q* is then rewritten:

$$Q(q_1, q_2, q_3, \ldots, q_n) = \left(\frac{1}{2\pi}\right)^{N/2} \exp\left(-\frac{1}{2}\chi^2\right)$$

χ^2 has a tabulated probability distribution which is usually written: $F(\chi^2)d\chi^2$.
The derivation of the expression for $F(\chi^2)$ is fairly abstract and complicated in most textbooks on the subject, but it can be done relatively simple if you consider: $\chi^2=\sum_{i=1}^{N} q_i^2$ as the distance to a point in an *N*-dimensional space.

The volume of a spherical shell with radius *r* in an *N*-dimensional space must necessarily be proportional to r^{N-1}.
In the plane the volume element is in polar coordinates: $dV_2 = rdrd\varphi$, and in 3–dim space it is: $dV_3 = r^2 sin\theta\, dr\, d\theta\, d\varphi$.
More generally if you calculate the Jacoby determinant for transforming from *n* dimensional ordinary coordinates to polar coordinates all the coordinates x_i have a factor *r* times a function of the *N*-1 angles.
By the partial differentiation of the coordinates x_i, with respect to the polar coordinates, the *r* will disappear in exactly one column, but remain in all the *N* - 1 other columns, and since the determinant is the of products of a factor from each column, the result will be that all terms have a factor r^{N-1}.
We start out by determining (except for a constant factor) the contribution to $F(\chi^2)d\chi^2$ from a spherical shell between the radii χ and χ + dχ and where χ is constant on the shell.

$$F(\chi^2)d\chi^2 = F(\chi^2)2\chi d\chi = \int_{\substack{shell\\ \chi,\chi+d\chi}} Q(q_1,q_2..q_n)dq_1dq_2...dq_n = \left(\frac{1}{2\pi}\right)^{N/2} \exp\left(-\frac{1}{2}\chi^2\right) \int_{\substack{shell\\ \chi,\chi+d\chi}} dq_1dq_2...dq_n$$

After the integration over the $N-1$ angles have been carried out, the integrand is, (according to the arguments stated above), proportional to the radius χ of the sphere to the $N-1$ power, which is χ^{N-1}.
If we collect the integral over the angles in a constant $2C$, we find:

$$F(\chi^2)d\chi^2 = F(\chi^2)2\chi d\chi = (2C)\chi^{N-1}\exp\left(-\frac{1}{2}\chi^2\right)d\chi$$
$$= C\chi^{N-2}\exp\left(-\frac{1}{2}\chi^2\right)2\chi d\chi = C\chi^{N-2}\exp\left(-\frac{1}{2}\chi^2\right)d\chi^2$$

The constant *C* is then determined by the normalization condition for a probability density function:

$$\int_0^\infty F(\chi^2)d\chi^2 = 1$$

The Gamma function is defined by the integral:

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt\ .$$

It is straight forward to derive the functional equation for the Gamma function: $\Gamma(x+1) = x\,\Gamma(x)$, since

$$\Gamma(x+1)=\int_0^\infty t^x e^{-t}dt=-\int_0^\infty t^x de^{-t}=-\left[t^x e^{-t}\right]_0^\infty+x\int_0^\infty t^{x-1}e^{-t}dt=x\Gamma(x)$$

And since $\Gamma(1)=1$, we have $\Gamma(2)=\Gamma(1+1)=1\cdot\Gamma(1)=1$, $\Gamma(3)=\Gamma(2+1)=2\cdot\Gamma(1)=2\cdot1=2!$, and so on. It then follows that for any integer n, that $\Gamma(n+1)=n!$

The normalization condition for a probability density function can therefore be expressed with the help of the Gamma function, making the substitution: $t=\frac{1}{2}\chi^2\Rightarrow\chi=\sqrt{2t}$ *and* $dt=\chi d\chi$ in the expression for $\int_0^\infty F(\chi^2)d\chi^2$. So we get:

$$\int_0^\infty F(\chi^2)d\chi^2=2C\int_0^\infty \chi^{N-1}\exp\left(-\frac{1}{2}\chi^2\right)d\chi=2C\int_0^\infty (2t)^{\frac{N-1}{2}}\exp(-t)\frac{1}{\sqrt{2t}}dt$$

$$=C\cdot 2^{\frac{N}{2}}\int_0^\infty t^{\frac{N}{2}-1}\exp(-t)dt=C\cdot 2^{\frac{N}{2}}\Gamma(\tfrac{N}{2})$$

So $C=(2^{\frac{N}{2}}\Gamma(\tfrac{N}{2})=1)^{-1}$. And the final expression for $F(\chi^2)d\chi^2$ becomes:

$$F(\chi^2)d\chi^2=\frac{1}{2^{\frac{N}{2}}\Gamma(\frac{N}{2})}e^{-\frac{1}{2}\chi^2}(\chi^2)^{\frac{N}{2}-1}d\chi^2$$

7. Pearson's formula for the χ^2 test

In high school and university text books relating to social sciences and biology there appear a formula, described as a χ^2 test, but it does not look like the formula based on the normal distribution derived above.

If we have a measurement of a quantity, then we denote the observed values of O by O_i and the expected values of O by E_i , Karl Pearson launched a test formula for the validity of a hypothesis to be used especially in social sciences and biology.

Although Pearson's formula looks different, it uses the χ^2 probability distribution as if the test was an ordinary χ^2 test based on the normal distribution.

Another strange thing about Pearson's formula is that it seemingly may be used, having only one observation. The formula reads:

$$\chi^2=\sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i}$$

This is in fact rather different from the usual formula for χ^2.

$$\chi^2=\sum_{i=1}^{n}\frac{(x_i-y_i)^2}{2\sigma_i^2}$$

Where x_i are observations having the theoretical value y_i and the standard deviation σ_i .

Pearson's formula has widespread applications, but it is most often applied to decide whether two quantities are correlated (dependent on each other)

If two statistical quantities are independent of each other, then you can find the probability that they both occur by multiplying the probabilities of their occurrence.

As usual $P(H)$ is the probability that an event occur in the outcome space U.

In a finite probability field, n or $n(U)$ is the number of elements in the outcome space U, and $n(H)$ is the number of elements in the event H.

The probability $P(H)$ correspond in statistics to the frequency of an observation $f(H) = \frac{n(H)}{n(U)}$

The two events A and B are independent if: $P(A \cap B) = P(A)P(B)$

The conditional probability that A occurs, if B is given, is written as: $P(A \mid B)$

If the events A and B are not independent the probability $P(A \cap B)$ should be calculated from the formula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \Leftrightarrow \quad P(A \cap B) = P(A \mid B)P(B)\,.$$

We shall now give a specific example (without referring to actual statistical material) on how one may examine whether two observations are dependent on each other or not.
So we shall examine the hypothesis:

Voting for Labour (S) is independent of whether you are an unskilled worker (W)or not.

The number of voters examined is $n = n(U)$. The numbers who vote for Labour is $n(S)$, the number who do not vote for Labour is $n(\overline{S})$, and such forth.
For example $n(W \cap \overline{S})$ is the number of unskilled workers that do not vote for Labour.
We therefore have 4 possibilities:

$$n(W \cap S),\ \ n(W \cap \overline{S}),\ n(\overline{W} \cap S),\ n(\overline{W} \cap \overline{S})\ .$$

These 4 possibilities are however not independent, as there exists two relations:

$$n(W \cap S) + n(W \cap \overline{S}) = n(U) \quad \text{and} \quad n(W \cap S) + n(\overline{W} \cap S) = n(S)$$

So as for S and for W we have 2 – 1 degrees of freedom, which account to (2-1)(2-1) = 1 degree of freedom.
Our hypothesis is that, when you vote S, then it is independent of whether you are unskilled or not.

We therefore assume: $P(W \cap S) = P(W)P(S)$

Then we write down Pearson's formula:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Which in this case has four terms.

$$\chi^2 = \frac{(n(W \cap S) - n(W)n(S))^2}{n(W)n(S))} + \frac{(n(\overline{W} \cap S) - n(\overline{W})n(S))^2}{n(\overline{W})n(S))} + \frac{(n(W \cap \bar{S}) - n(W)n(\bar{S}))^2}{n(W)n(\bar{S}))} + \frac{(n(\overline{W} \cap \bar{S}) - n(\overline{W})n(\bar{S}))^2}{n(\overline{W})n(\bar{S}))}$$

If we insert factual numbers – for example from the last election to the Danish parliament, we shall obtain a value for χ^2.

If we get a number on 3.84, then the probability of getting a greater value than the calculated one is about 0.95. The value therefore belongs to the 95% accept interval, and the hypothesis may be accepted, but if we get a value 10 times larger, when evaluating χ^2, then the probability of getting a higher value is 0.00157.
This value is so small, that it is highly improbable, that it is due to statistical fluctuations, and we should discard the hypothesis:
Voting for Labour is independent of whether you are unskilled or not.

Using real numbers the significance of this test, is probably even larger than is this constructed case.

8. Explaining Pearson's formula.

At a glance there is only little similarity between Pearson's formula and the χ^2 distribution test of a hypothesis. One reason is that whereas the χ^2 probability distribution is founded on the normal distribution, Pearson's formula is based on the binomial distribution.
We shall however try to justify that Pearson's formula may nevertheless be conceive as a χ^2 – test.

A stochastic variable *X*, is said to have a *binomial distribution*, if the probability that an event occurs is *p*, and where the probability that the event occurs *q* times in a series of *n* is given by

$$P(X = q) = \binom{n}{q} p^q (1-p)^{n-q}$$

The mean value is $\mu = np$ and the standard deviation is $\sigma = \sqrt{np(1-p)}$

We then write Pearson's formula

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

The stochastic variable X, which is assumed to be binomially distributed, and its complementary event $\overline{X}$, that is, $X \cap \overline{X} = \emptyset$ and $X \cup \overline{X} = U$ are not independent since they exclude each other and together represents the outcome space, the number of degrees of freedom is 2-1 = 1.
If the experiment is repeated *n* times and *X* occurs *q* times then $\overline{X}$ occurs *n-q* times.

Furthermore:

$$E(X) = np \quad \text{and} \quad \sigma(X) = \sqrt{np(1-p)} \quad \text{and} \quad E(\overline{X}) = n(1-p) \quad \textit{and} \quad \sigma(\overline{X}) = \sqrt{n(1-p)p}$$

According to Pearson's formula:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{(X_i - np)^2}{np} + \frac{((n - X_i) - n(1-p))^2}{n(1-p)} \right)$$

We shall settle for one of the terms in the sum dropping the index *i*, and submit the term to a series of rearrangements.

$$\chi_i^{\,2} = \frac{(X - np)^2}{np} + \frac{((n - X) - n(1-p))^2}{n(1-p)} = \frac{(X - np)^2}{np} + \frac{(np - X)^2}{n(1-p)} =$$

$$\frac{(1-p)(X - np)^2 + p(np - X)^2}{np(1-p)} = \frac{(X - np)^2}{np(1-p)} = \frac{(X - \mu)^2}{\sigma^2}$$

We have thus transformed (assuming a binomial distribution) Pearson's formula into the standard expression for the Chi-square expression: so

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad \text{is equivalent to} \quad \chi^2 = \sum_{k=1}^{n} \frac{(x_k - \mu_k)^2}{\sigma_k^{\,2}}$$

Which shows that that Pearson's formula is actually a χ^2 – test.

One could object that χ^2 refers to the normal distribution, while Pearson's formula is founded on the binomial distribution. However, we have shown that the binomial distribution asymptotically can be represented by the normal distribution the limit, where $n \to \infty$.

From this it follows that one may actually apply Person's formula as a χ^2 test, having one degree of freedom.