

# Statistics for EES and MEME

## Chi-square tests and Fisher's exact test

Dirk Metzler

April 7, 2026

### Contents

1	$X^2$ goodness-of-fit test	1
2	$X^2$ test for homogeneity/independence	3
3	Fisher's exact test	6
4	$X^2$ test for fitted models with free parameters	8

### 1 $X^2$ goodness-of-fit test

#### Mendel's experiments with peas

green (recessive) vs. yellow (dominant)

round (dominant) vs. wrinkled (recessive)

Expected frequencies when crossing double-hybrids:

	green	yellow
wrinkled	$\frac{1}{16}$	$\frac{3}{16}$
round	$\frac{3}{16}$	$\frac{9}{16}$

Observed in experiment ( $n = 556$ ):

	green	yellow
wrinkled	32	101
round	108	315

Do the observed frequencies agree with the expected ones?

Relative frequencies:

	green/wrink.	yell./wrink.	green/round	yell./round
expected	0.0625	0.1875	0.1875	0.5625
observed	0.0576	0.1942	0.1816	0.5665

Can these deviations be well explained by pure random?

Measure deviations by  $X^2$ -statistic:

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where  $E_i$  = expected number in class  $i$  and  $O_i$  = observed number in class  $i$ .  
 Why scaling  $(O_i - E_i)^2$  by dividing by  $E_i = \mathbb{E}O_i$ ?

Let  $n$  be the total sample size and  $p_i$  be the probability (under the null hypothesis) each individual to contribute  $O_i$ .

Under the null hypothesis,  $O_i$  is binomially distributed:

$$\Pr(O_i = k) = \binom{n}{k} p_i^k \cdot (1 - p_i)^{n-k}.$$

Thus,

$$\mathbb{E}(O_i - E_i)^2 = \text{Var}(O_i) = n \cdot p \cdot (1 - p).$$

If  $p$  is rather small,  $n \cdot p \cdot (1 - p) \approx n \cdot p$  and

$$\mathbb{E} \frac{(O_i - E_i)^2}{E_i} = \frac{\text{Var}(O_i)}{\mathbb{E}O_i} = 1 - p \approx 1.$$

By the way...

the binomial distribution with small  $p$  and large  $n$  can be approximated by the Poisson distribution:

$$\binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k} \approx \frac{\lambda^k}{k!} \cdot e^{-\lambda} \quad \text{with} \quad \lambda = n \cdot p.$$

A random variable  $Y$  with possible values  $0, 1, 2, \dots$  is *Poisson distributed* with parameter  $\lambda$ , if

$$\Pr(Y = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}.$$

Then,  $\mathbb{E}Y = \text{Var}(Y) = \lambda$ .

	g/w	y/w	g/r	y/r	sum
theory	0.0625	0.1875	0.1875	0.5625	
expected	34.75	104.25	104.25	312.75	556
observed	32	101	108	315	556
$O - E$	-2.75	-3.25	3.75	2.25	
$(O - E)^2$	7.56	10.56	14.06	5.06	
$\frac{(O-E)^2}{E}$	0.22	0.10	0.13	0.02	0.47

$$X^2 = 0.47$$

Is a value of  $X^2 = 0.47$  usual?

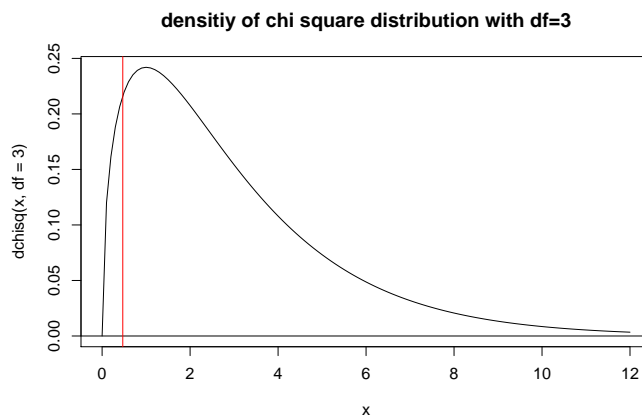
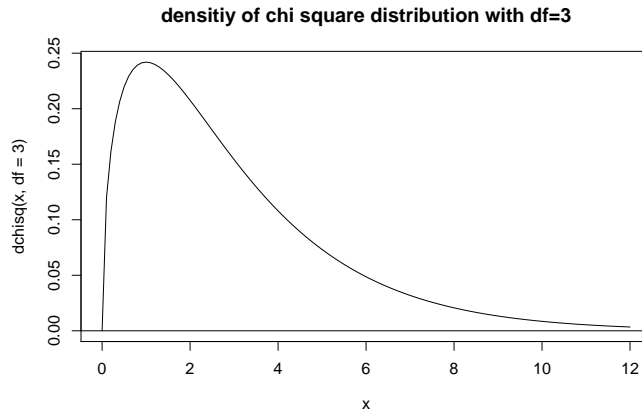
The distribution of  $X^2$  depends on the *degrees of freedom* (df).

in this case: the sum of the observations must be  $n = 556$ .

↪ when the first three numbers 32, 101, 108 are given, the last one is determined by

$$315 = 556 - 32 - 101 - 108.$$

$$\Rightarrow df = 3$$



```
> pchisq(0.47,df=3)[1ex] [1] 0.07456892
```

*p*-value = 92.5%

```
> obs <- c(32,101,108,315)
> prob <- c(0.0625,0.1875,0.1875,0.5625)
> chisq.test(obs,p=prob)
```

Chi-squared test for given probabilities

```
data: obs
X-squared = 0.47, df = 3, p-value = 0.9254
```

## 2 $X^2$ test for homogeneity/independence

The cowbird is a brood parasite of Oropendola



[http://commons.wikimedia.org/wiki/File:Montezuma\\_Oropendola.jpg](http://commons.wikimedia.org/wiki/File:Montezuma_Oropendola.jpg) photo (c) by J. Oldenettel

## References

[Smi68] N.G. Smith (1968) The advantage of being parasitized. *Nature*, **219(5155)**:690-4

- Cowbird eggs look very similar to oropendola eggs.
- Usually, oropendola rigorously remove all eggs that are not very similar to theirs.
- In some areas, cowbird eggs are quite different from oropendola eggs but are tolerated.
- Why?
- Possible explanation: botfly (german: Dasselfliegen) larvae often kill juvenile oropendola.
- nests with cowbird eggs are somehow better protected against the botfly.

	no. of cowbird eggs	0	1	2
numbers of nests affected by botflies	affected by botflies	16	2	1
	not affected by botflies	2	11	16

	no. of cowbird eggs	0	1	2
percentages of nests affected by botflies	affected by botflies	89%	15%	6%
	not affected by botflies	11%	85%	94%

- apparently, the infestation with botflies is reduced when the nest contains cowbird eggs
- statistically significant?
- null hypothesis: The probability of a nest to be affected with botflies is independent of the presence of cowbird eggs.

	no. of cowbird eggs	0	1	2	$\Sigma$
numbers of nests affected by botflies	affected by botflies	16	2	1	19
	not affected by botflies	2	11	16	29
	$\Sigma$	18	13	17	48

which numbers of affected nests would we expect under the null hypothesis?

The same ratio of 19/48 in each group.

expected numbers of nests affected by botflies, given row sums and column sums

no. of cowbird eggs	0	1	2	$\Sigma$
affected by botflies	7.1	5.1	6.7	19
not affected by botflies	10.9	7.9	10.3	29
$\Sigma$	18	13	17	48

$$18 \cdot \frac{19}{48} = 7.125 \quad 13 \cdot \frac{19}{48} = 5.146$$

All other values are now determined by the **sums**. (caution: rounding errors!)

Observed (O):	affected by botflies	16	2	1	19
	not affected by botflies	2	11	16	29
	$\Sigma$	18	13	17	48

Expected (E):	affected by botflies	7.1	5.1	6.7	19
	not affected by botflies	10.9	7.9	10.3	29
	$\Sigma$	18	13	17	48

O-E:	affected by botflies	8.9	-3.1	-5.7	0
	not affected by botflies	-8.9	3.1	5.7	0
	$\Sigma$	0	0	0	0

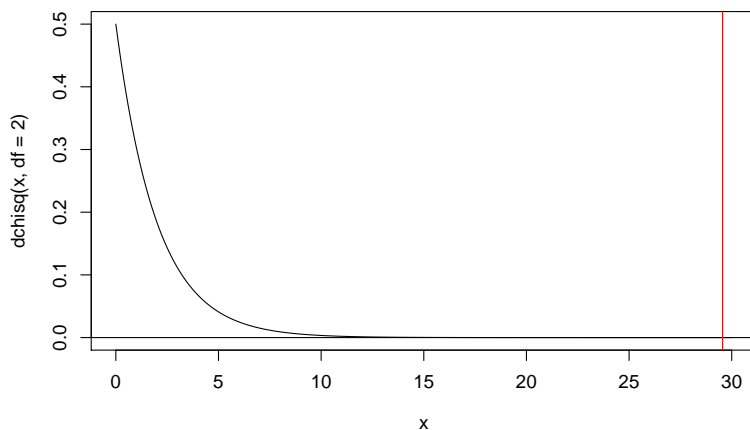
(more precisely:  $8.875 - 3.145833 - 5.729167 = 0$ )

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = 29.5544$$

- given the sums of rows and columns, two values in the table determine the rest
- $\Rightarrow$   $df=2$  for contingency table with 2 rows and 3 columns
- in general for tables with  $n$  rows and  $m$  columns:

$$df = (n - 1) \cdot (m - 1)$$

**density of chi square distribution with  $df=2$**



```

> M <- matrix(c(16,2,2,11,1,16),nrow=2)
> M
      [,1] [,2] [,3]
[1,]   16    2    1
[2,]    2   11   16
> chisq.test(M)

Pearson's Chi-squared test

data:  M
X-squared = 29.5544, df = 2, p-value = 3.823e-07

The p-value is based on approximation by  $\chi^2$ -distribution.
Rule of thumb:  $\chi^2$ -approximation appropriate if all expected values are  $\geq 5$ .
Alternative: approximate p-value by simulation:
> chisq.test(M,simulate.p.value=TRUE,B=50000)

Pearson's Chi-squared test with simulated p-value
(based on 50000 replicates)

data:  M
X-squared = 29.5544, df = NA, p-value = 2e-05

```

### 3 Fisher's exact test

## References

[McK91] J.H. McDonald, M. Kreitman (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**:652-654.

	synonymous	replacement
polymorphisms	43	2
fixed	17	7

```

> McK <- matrix(c(43,17,2,7),2,
               dimnames=list(c("polymorph","fixed"),
                             c("synon","replace")))
> McK
      synon replace
polymorph  43     2
fixed      17     7
> chisq.test(McK)

Pearson's Chi-squared test
with Yates' continuity correction

data:  McK
X-squared = 6.3955, df = 1, p-value = 0.01144

Warning message: In chisq.test(McK) :
Chi-Square-Approximation may be incorrect

```

```
> chisq.test(McK,simulate.p.value=TRUE,B=100000)
```

Pearson's Chi-squared test with simulated p-value  
(based on 1e+05 replicates)

data: McK

X-squared = 8.4344, df = NA, p-value = 0.00649

### Fisher's exact test

$$\begin{array}{c|c} A & B \\ \hline C & D \end{array}$$

- null hypothesis:  $\frac{\mathbb{E}A/\mathbb{E}C}{\mathbb{E}B/\mathbb{E}D} = 1$
- For  $2 \times 2$  tables **exact**  $p$ -values can be computed (no approximation, no simulation).

```
> fisher.test(McK)
```

### Fisher's Exact Test for Count Data

data: McK

p-value = 0.006653

alternative hypothesis: true odds ratio  
is not equal to 1

95 percent confidence interval:

1.437432 92.388001

sample estimates:

odds ratio

8.540913

$$\begin{array}{cc|c} 43 & 2 & \sum \\ \hline 17 & 7 & 24 \\ \hline \sum & 60 & 9 & 69 \end{array} \qquad \begin{array}{cc|c} a & b & \sum \\ \hline c & d & M \\ \hline \sum & U & V & N \end{array}$$

Given the row sums and column sums and assuming independence, the probability of  $a$  is

$$\Pr(a) = \frac{\binom{K}{a}\binom{M}{c}}{\binom{N}{U}} = \Pr(b) = \frac{\binom{K}{b}\binom{M}{d}}{\binom{N}{V}}$$

“hypergeometric distribution”

$p$ -value:

$$\Pr(b = 0) + \Pr(b = 1) + \Pr(b = 2)$$

	a	b	$\Sigma$
	c	d	45
$\Sigma$	60	9	24
			69
$b$	$\Pr(b)$		
0	0.000023		
1	0.00058		
2	0.00604		
3	0.0337		
4	0.1117		
5	0.2291		
6	0.2909		
7	0.2210		
8	0.0913		
9	0.0156		

**One-sided Fisher test:**

for  $b = 2$ :  
 $p\text{-value} = \Pr(0) + \Pr(1) + \Pr(2) = 0.00665313$   
for  $b = 3$ :  
 $p\text{-value} = \Pr(0) + \Pr(1) + \Pr(2) + \Pr(3) = 0.04035434$

**Two-sided Fisher test:**

Sum up all probabilities that are smaller or equal to  $\Pr(b)$ .  
for  $b = 2$ :  
 $p\text{-value} = \Pr(0) + \Pr(1) + \Pr(2) = 0.00665313$   
for  $b = 3$ :  
 $p\text{-value} = \Pr(0) + \Pr(1) + \Pr(2) + \Pr(3) + \Pr(9) = 0.05599102$

## 4 $\chi^2$ test for fitted models with free parameters

Given a population in *Hardy-Weinberg equilibrium* and a gene locus with two alleles A and B with frequencies  $p$  and  $1 - p$ .

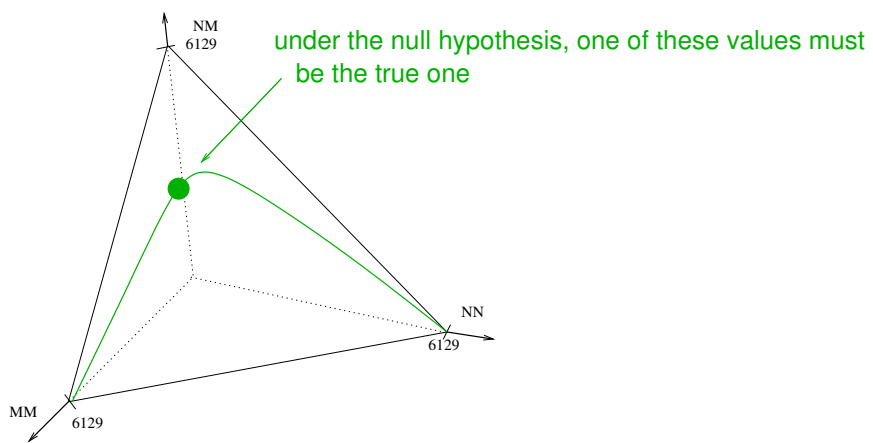
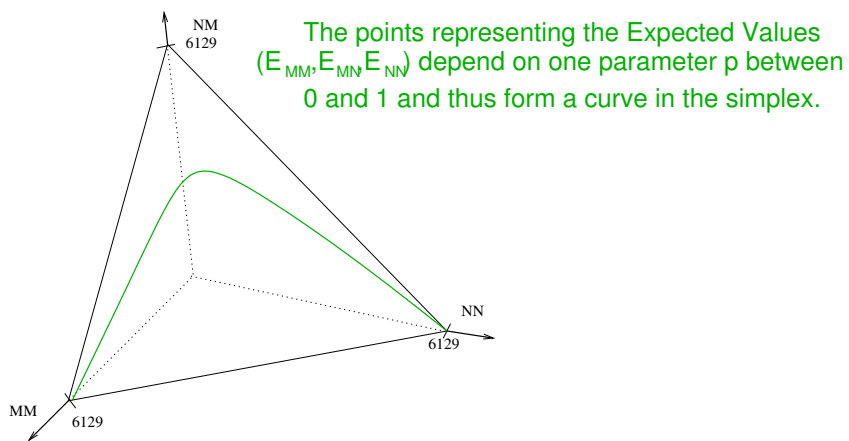
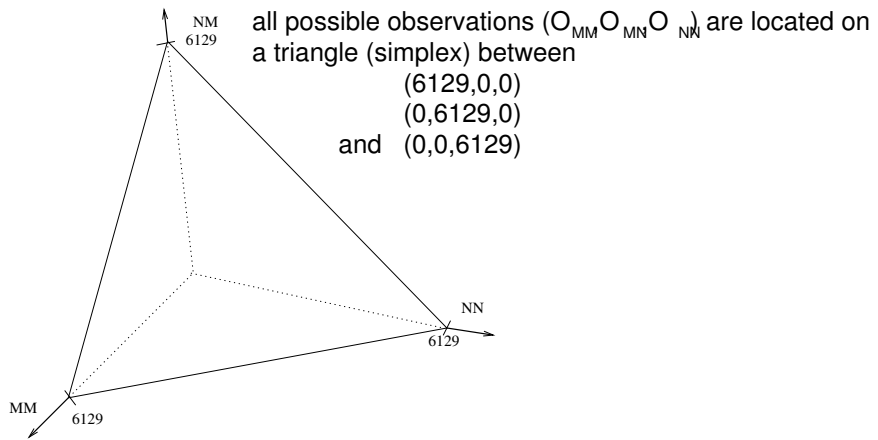
↪ Genotype frequencies

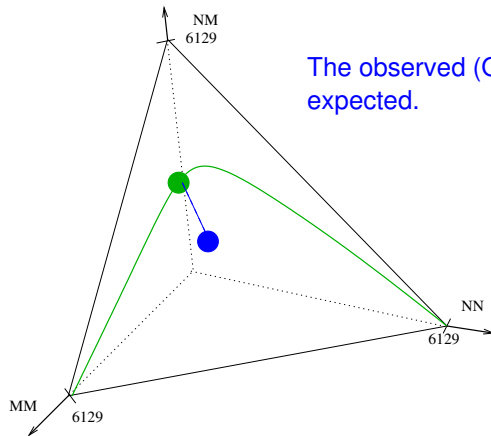
AA	AB	BB
$p^2$	$2 \cdot p \cdot (1 - p)$	$(1 - p)^2$
example: M/N blood type;	sample: 6129 white Americans	
observed: MM	MN	NN
1787	3037	1305

estimated allele frequency  $p$  of M:

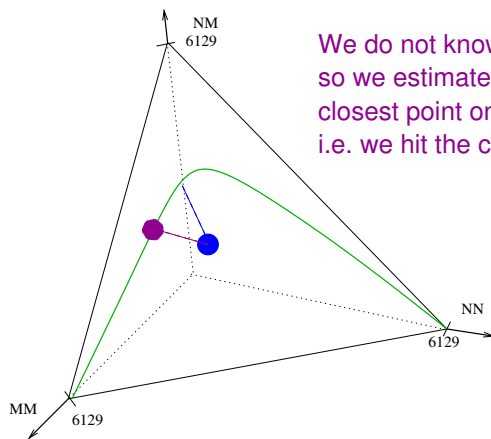
$$\frac{2 \cdot 1787 + 3037}{2 \cdot 6129} = 0.5393$$

↪ expected:	MM	MN	NN
	$p^2$	$2 \cdot p \cdot (1 - p)$	$(1 - p)^2$
	0.291	0.497	0.212
	1782.7	3045.5	1300.7

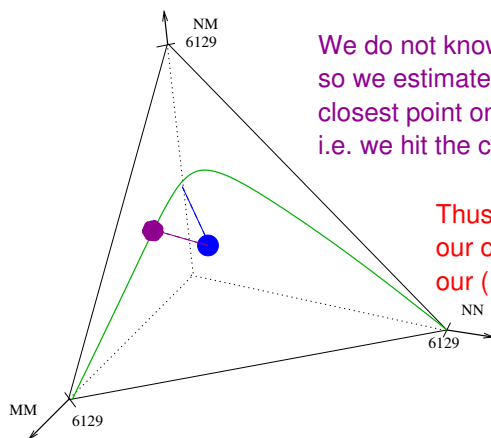




The observed  $(O_{MM}, O_{NM}, O_{NN})$  will deviate from the expected.



We do not know the true expectation values so we estimate  $(E_{MM}, E_{NM}, E_{NN})$  by taking the closest point on the curve of possible values, i.e. we hit the curve in a right angle.



We do not know the true expectation values so we estimate  $(E_{MM}, E_{NM}, E_{NN})$  by taking the closest point on the curve of possible values, i.e. we hit the curve in a right angle.

Thus, deviations between our observations  $(O_{MM}, O_{NM}, O_{NN})$  and our  $(E_{MM}, E_{NM}, E_{NN})$  can only be in one dimension: perpendicular to the curve.

$$df = k - 1 - m$$

$k$  = number of categories ( $k=3$  genotypes)  $m$  = number of model parameters ( $m=1$  parameter  $p$ ) in blood

type example:

$$df = 3 - 1 - 1 = 1$$

```
> p <- (2* 1787+3037)/(2* 6129)
> probs <- c(p^2,2*p*(1-p),(1-p)^2)
> X <- chisq.test(c(1787,3037,1305),p=probs)$statistic[[1]]
> p.value <- pchisq(X,df=1,lower.tail=FALSE)
> X
[1] 0.04827274
> p.value
[1] 0.8260966
```

**Test result:** According to the chi-square test the data show no significant deviation from a Hardy-Weinberg equilibrium ( $X^2 = 0.048$ ,  $df = 1$ ,  $p = 0.83$ ).

**Wrong would be:** “We conclude that the population is in Hardy-Weinberg equilibrium (for this gene locus).”

**Reason:** Statistical tests can never show that a null hypothesis is fulfilled.

### Some of what you should be able to explain

- $X^2$ -statistic: structure and idea
- df of different variants of  $X^2$  test
- $\chi^2$  distributions: when and how to use them
- Fisher’s exact test
  - When applicable?
  - hypergeometric distribution
  - How, exactly, to apply two-sided
- Hardy-Weinberg equilibrium