STATISTICS IN EVALUATING THE MEIOSIS OF CEREALS

1. INTRODUCTION: MEIOSIS, CROSSOVER AND INTERFERENCE

Meiosis is the cellular division that takes place during sexual reproduction, allowing the production of germ cells. During meiosis the phenomenon of crossover, also known as genetic recombination, introduces genetic variability through the exchange of genetic material between homologous chromosomes. The meiotic division therefore allows individuals that reproduce sexually to have a genetically diverse offspring, which has several advantages regarding evolution, such as a better adaptation to the environment.

Meiosis involves two sequential divisions, creating four haploid (n) cells from a single diploid (2n) cell. During the prophase of the first meiotic division, the crossover occurs between homologous chromosomes, that interchange several segments.

Even though the crossover events were initially thought to be independent, it is widely known that some genes are linked. Between loci of linked genes, the event of a recombination is low likely to happen.

On the other hand, most of the genes are inherited independently, and are either located in different chromosomes or at a certain genetic distance (>0.5 centiMorgans, cM) in the same chromosome. A distance over 0.5 cM means that a crossover event takes place many times between the two loci.

Additionally, it is known that crossovers are not distributed evenly throughout the chromosome, and interference may occur.

In the absence of interference, crossovers would occur independently of each other, and the location of the crossover on a chromosome would be randomly distributed. However, the existence of crossover interference suggests that the occurrence of one crossover event can influence the likelihood or probability of a nearby crossover event.

There are two types of interference:

A. Positive interference: it is less likely that another crossover will occur nearby. This means that the presence of a crossover inhibits or interferes with the occurrence of additional crossovers in the same region.
B. Negative interference: the occurrence of a crossover in one region increases the likelihood of a crossover occurring nearby. In this case, the presence of a crossover promotes additional crossovers in its vicinity.

Understanding the distribution of recombination events along a chromosome, as well as the crossover interference is important in genetics, as it has implications for the inheritance patterns of genetic traits in individuals that have a sexual reproduction, such as cereals.

In this essay, the inheritance of a group of morphological characteristics will be evaluated for the Oregon Wolf Barley (OWB), to illustrate with an example the use of the chi-square goodness-of-fit test in genetics. The use of statistic models to evaluate the crossover in barley and other cereals will also be discussed.

1. CHI-SQUARE GOODNESS-OF-FIT TO DETERMINE THE INDEPENDECE OF GENES

1.1. THEORETICAL BACKGROUND

The common use of the chi-square (χ^2) model in the context of genomic maps is as a goodness-of-fit test between the observed data and the expected distribution of crossovers.

First, the null hypothesis is defined:

- Null Hypothesis (H_0): Assumes no interference. The observed distribution of crossovers follows the expected distribution.
- Alternative Hypothesis (H_a): Assumes interference. The observed distribution deviates significantly from the expected distribution.

Then, a calculation of the expected distribution of crossovers under the assumption of the null hypothesis being true is done:

Where:

- Oi is the observed frequency.
- Ei is the expected frequency.

The degrees of freedom (df) are determined by the number of categories minus 1.

The next step is to compare the calculated chi-square statistic with the critical value from the chi-square distribution table. By searching in the table, and depending on the degrees of freedom, a p value can be obtained. A statistical software can also be used to obtain the p-value. Once obtained the p-value we can evaluate whether to reject or not the null hypothesis:

A. If the p-value is less than the chosen significance level (α), reject the null hypothesis. This suggests the presence of crossover interference.
B. If the p-value is greater than α, the null hypothesis is not rejected, indicating no significant deviation from the expected distribution.

1.2. APLICATION OF χ^2 GOODNESS-OF-FIT TO THE CALCULUS OF INHERITANCE OF MORPHOLOGICAL CHARACTERS IN CEREALS

This test allows us to perform a genetic map from observed data. To illustrate this with an example, we will use the data of OWB.

The first step is knowing which individual segregation has each character. We will assume that the characters follow a dominant mendelian segregation (3:1).

- H_0 = The observed proportion is 3:1, meaning that the character has a dominant mendelian segregation
- H_a= The observed proportion is not 3:1

Number of rows		
	2	6
Observed	167	43
Expected (3:1)	157,5	52,5
χ^2	2,2921	
Grain		
	Dressed	Naked
Observed	166	44
Expected (3:1)	157,5	52,5
χ^2	1,8349	
Variegation		
	Yes	No
Observed	48	162
Expected (3:1)	52,5	157,5
χ^2	0,51429	

Awn type		
	Normal	Trifurcated
Observed	102	108
Expected (3:1)	52,5	157,5
χ^2	62,223	
Type of spike		
	Dense	Normal
Observed	154	56
Expected (3:1)	157,5	52,5
χ^2	0,31111	
Leaf pubescence		
	Yes	No
Observed	137	45
Expected (3:1)	136,5	45,5
χ^2	0,0073260	

As there is 1 degree of freedom, we search the χ^2 in the reference table and obtain that the awn type does not follow the (3:1) distribution traits ($\chi^2 > 3{,}841$), so we reject the null hypothesis for this character.

In the case of number of rows, grain, variegation, leaf pubescence and type of spike, the null hypothesis cannot be rejected.

There are other distributions for the F1 of a single character, with different segregation patterns:

- Simple recessive epistasis 9:3:4
- Simple dominant epistasis 12:3:1

- Double recessive epistasis 9:7
- Double dominant epistasis 15:1
- Double dominant-recessive epistasis 13:3

It is feasible to think that the double recessive epistasis could fit with the observed data, so a new null hypothesis is defined:

- H_0= The proportion is 9:7, meaning that the character has a double recessive epistasis inheritance.
- H_a= The proportion is not 9:7, so the character has another type of inheritance

By performing the same analysis as seen before, with a contingency table, the value of χ^2 obtained is 1,9837 (<3,841 for 1 degree of freedom), so this time the null hypothesis is not rejected.

The conclusion of this analysis is that the characteristic "awn type" is defined by more than one gene, whose locations are unknown. It is important to take this into account when building the genetic map and analysing the linkage to other genes.

The following step is to analyse whether the barley morphological characters are inherited independently. The expected inheritance pattern would be the classic F2 (9:3:3:1).

- H_0= The morphological treats follow a 9:3:3:1 inheritance, meaning that they are independent
- H_a= The morphological treats do not follow a 9:3:3:1 inheritance, meaning that they are linked

A contingency table and an analysis of fitness-of-fit with χ^2 is performed for all the combinations:

Grain					
		Dressed		Naked	
		Observed	Expected	Observed	Expected
Variegation	No	124	127,29	38	34,71
	Yes	41	37,71	7	10,29
$\chi^2 = 1,7317$					
Leaf pubescence	Yes	112	109,15	25	27,85
	No	33	35,85	12	9,15
$\chi^2 = 1,4822$					
Number of rows	2	126	129,64	39	35,36
	6	39	35,36	6	9,64
$\chi^2 = 2,2292$					
Awn type	Normal	52	74,54	113	90,36
	Trifurcated	43	20,36	2	24,64
$\chi^2 = 58,533$					
Type of spike	Dense	117	114,71	29	31,29
	Normal	48	50,29	16	13,71
$\chi^2 = 0,69738$					
Variegation					
		No		Yes	
		Observed	Expected	Observed	Expected
Leaf pubescence	Yes	100	103,13	37	33,87
	No	37	33,87	8	11,13
$\chi^2 = 1,5503$					
Number of rows	2	35	37,71	13	10,29
	6	130	127,29	32	34,71
$\chi^2 = 1,1817$					
Awn type	Normal	19	21,71	76	73,29
	Trifurcated	29	26,29	86	88,71
$\chi^2 = 0,8031$					
Type of spike	Dense	20	33,37	126	112,63
	Normal	28	14,63	36	49,37
$\chi^2 = 22,789$					

Leaf pubescence					
		Yes		No	
		Observed	Expected	Observed	Expected
Number of rows	2	104	106,89	33	30,11
	6	33	35,11	7	9,89
$\chi^2 = 1,4380$					
Awn type	Normal	58	61,73	24	20,27
	Trifurcated	79	75,27	21	24,73
$\chi^2 = 1,6549$					
Type of spike	Dense	96	95,60	31	31,40
	Normal	41	41,40	14	13,60
$\chi^2 = 0,022522$					
Number of rows					
		2		6	
		Observed	Expected	Observed	Expected
Awn type	Normal	78	74,64	17	20,36
	Trifurcated	87	90,36	28	24,64
$\chi^2 = 1,2867$					
Type of spike	Dense	122	114,7	21	31,29
	Normal	43	50,29	24	13,71
$\chi^2 = 7,0855$					
Awn type					
		Dressed		Naked	
		Observed	Expected	Observed	Expected
Type of spike	Dense	63	66,05	83	79,95
	Normal	32	28,95	32	35,05
$\chi^2 = 0,84261$					

The analysis of joint segregation indicates the existence of linkage between the following morphological traits. $\chi2 > 3,8415$ indicates rejection of the null hypothesis for 1 degree of freedom (calculated as (n-1) x (m-1) = 1 x 1 = 1) with a confidence of 95%. The rejection of the null hypothesis would mean that the recombination fraction, r, is lower than 0.5 and the genes are linked:

- Grain and awn type
- Rows and spike type (r: 0.38723)
- Spike type and variegation (r: 0.3042)

All of them are linked in coupling phase. The estimation of the recombination fraction employs the formula corresponding to an F2 with dominance at both loci, that is:

$$nx^2+(-a_1+2a_2+2a_3+a_4)x-2a_4$$

$$x=r^2 \text{ (repulsion phase)}$$

$$x=(1-r)^2 \text{ (coupling phase)}$$

In repulsion phase, each homologous chromosome carries one dominant allele and one recessive allele, whereas in coupling phase each homologous chromosome carries both dominant alleles or both recessive alleles. The phases are determined by the occurrence or absence of crossover events during the meiosis.

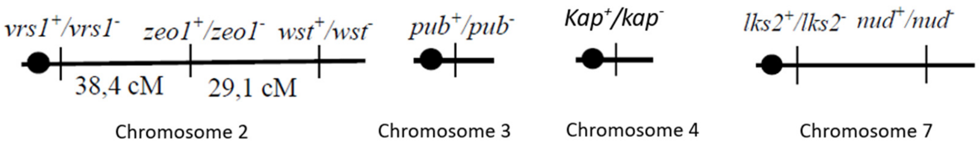
Because the awn type is not following a dominant inheritance, the r cannot be calculated with the provided formula (there is more than one locus defining this character). A recombination fraction below 0.5 indicates that the loci are linked, and that there is less probability that a crossover takes place between both loci.

The genetic distance (d) is the product of r times 100. So:

- Number of rows and spike type: d=38,723 cM
- Spike type and variegation: d=30,42 cM

1.3. Building of the genetic map

With the processed data, it is possible to elaborate a genetic map that includes some of the morphological characteristics and their location in the chromosome (OWB):



Where:

- Vrs1^{+}/vrs1^{-} : number of rows
- Zeo1^{+}/zeo1^{-} : type of spike
- Wst^{+}/wst^{-} : variegation
- Pub^{+}/pub^{-} : leaf pubescence
- Kap^{+}/kap^{-} and lks2^{+}/lks2^{-} : awn type
- Nud^{+}/nud^{-} : grain

2. STATISTICAL MODELS TO EVALUATE CROSSOVERS IN A CHROMOSOME

2.1. Count-Location (CL) model

In this model, the number of chiasma in four-strand bundles follows a certain distribution, denoted as p=(p_0, p_1, p_2…) and the locations of these chiasmata, given their number, are determined by placing them uniformly at random.

p_0, p_1, p_2, …, p_n are the probabilities of having 0, 1, 2 and n chiasmata, respectively.

The count-location model assumes that chiasma locations are independent and evenly distributed. Each possible location of the chiasmata is equally likely, and the occurrence and location of one chiasma are independent of the occurrence and location of another. This independence provides a simplified probabilistic description but does not consider the existence of interference. (Broman, K. W., & Weber, J. L., 2000).

Several special cases of this model are important to mention:

A. The "obligate-chiasma CL model", where p0 = 0, which means that there must be at least one chiasma on 0 of the four-strand bundle.
B. The no-interference model. In this case, we assume the distribution p is a Poisson distribution, so $P_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, where x stands for the number of chiasmata on a four-strand bundle, and λ would represent the average rate of chiasmata occurrence.
C. The "truncated Poisson model" considers both p0=0 and for x>0. In this case the model has an obligate chiasma, but in which there otherwise is no interference.

In the case of the CL model, including the mentioned special cases, the expected crossovers (EC) can be calculated, and are given by the following formula:

$$EC = \frac{1}{2} \cdot (1 - e^{-2 \cdot r})$$

Where r stands for the recombination fraction.

Regarding the example of the OWB, it is possible to calculate the number of expected crossovers between the number of rows and type of spike, as well as between variegation and type of spike.

$$EC_{number\ rows-type\ spike} = \frac{1}{2} \cdot (1 - e^{-2 \cdot 0{,}38723}) = 0{,}270$$

$$EC_{variegation-type\ spike} = \frac{1}{2} \cdot (1 - e^{-2 \cdot 0{,}3042}) = 0{,}228$$

However, it should be noted that this is not the only use of the CL model. By providing expected crossover frequencies, the count-location model helps estimate genetic distances in cM between genetic loci. It is also used for genetic mapping, quantification of recombination, analysis of linkage and to study crossover interference. Nonetheless, there are other models that assess crossover interference with more precision, such as the gamma model.

2.2. Gamma model

The gamma model of crossover interference is a statistical model used to describe the distribution of crossovers along a chromosome.

The gamma model assumes that crossovers are not randomly distributed along a chromosome. Instead, it suggests that the probability of a crossover occurring at a specific location is influenced by the presence of other crossovers in nearby regions, introducing in this way the concept of interference (Broman, K. W., & Weber, J. L., 2000)

In the gamma model, the probability of a crossover to take place at a certain position of the chromosome follows the gamma distribution. The probability density distribution is defined as:

$$f(x; k, \theta) = \frac{x^{k-1} e^{-x/\theta}}{\theta^k \Gamma(k)}$$

Where:

- x: crossover distance.
- k: shape parameter.
- θ: scale parameter.
- Γ(k): gamma function.

It describes the probability of observing a crossover at a specific distance X along the chromosome.

This gamma distribution has two parameters, and it can model situations in which crossovers are more clustered or dispersed throughout the chromosome.

The shape parameter (k) has a key role, as it determines the strength of the interference. Higher k values suggest stronger interference and crossovers evenly distributed, whereas lower k values will imply weaker interference, which means that it will approach a more random pattern.

Genetic distance (in Morgans) is the expected number of crossover events between two points on the genetic map. In terms of cM distances: E(D)=100=vθ. Hence, θ=100/v and to specify the model we need only parameter k. The case with k=1 corresponds to the situation of no crossover interference, k>1 implies positive COI, and 0<k<1 is negative interference (Sapielkin, S. *et al* 2022).

In this case, the expected number of crossovers within a specified region is given by the integral of the probability density function over that region, even though the formula expressed in the CL model can also be used, once estimated the appropriated parameters.

$$EC[X] = \int_a^b x f(x; k; \theta) dx$$

The gamma model allows researchers to fit the observed distribution of crossovers to the expected gamma distribution, by adjusting the values of k and θ in a particular set of data.

3. CALCULATION OF THE CROSSOVER INTERFERENCE IN CEREALS

3.1. OBSERVED DATA – OWB

To evaluate the interference in the observed data of the OWB, the coefficient of coincidence (C) will be used. C is a measure used in genetics to quantify the degree of interference in the occurrence of crossovers during meiosis, by comparing the observed and expected frequencies of double crossovers.

C can be defined as the following:

$$C = \frac{Observed\ double\ crossovers}{Expected\ double\ crossovers}$$

Where the observed double crossovers are the actual instances where two crossovers occur in proximity during meiosis, and the expected double crossovers are those that would be expected in the case of no interference (if crossovers were occurring independently of each other).

In this analysis of OWB, the observed double crossovers are:

- n7: vrs1^{+}zeo1^{-} wst^{+} - 22
- n8: vrs1^{-} zeo1^{+} wst $^{-}$ - 3

With this data, C can be calculated:

$$C = \frac{(n_7 + n_8)/total\ observations}{r_1 \cdot r_2} = \frac{(22+3)/210}{0{,}387 \cdot 0{,}304} = 1{,}011$$

Now, interference can be expressed as:

$$I{=}1{-}C{=}1{-}1{,}011{=}-0{,}011$$

Which indicates the existence of negative interference. This is extremely rare, as it does not normally occur in meiosis, but may be observed during virus recombination. The gathering of more data and the use of statistical tools would clarify the results.

3.2. USE OF THE GAMMA MODEL TO CALCULATE CROSSOVER INTERFERENCE IN CEREALS

To applicate the gamma model to the example of OWB, the first thing to be done is to estimate the shape and scale parameters. This is done using Statistics specific programs, such as R. Nonetheless, the gamma model is particularly used to explain positive interference, and in the observed data a negative interference was obtained.

Also, it should be noted that using an F2 population (such as the OWB example) results in substantial loss of statistical power to detect interference. This is due to the existence of dependent meiotic events present in an F2 individual. Using an F2 population, together with the analysis of small intervals, results in an overestimation of double recombination events, high values for the coefficient of coincidence (C >1) which can lead to measuring negative interference (Saintenac, C *et al*. 2009).

This statistical model has been used to study interference in barley. In 2016, Isabelle Colas and colleagues published a study in which they recollected data of MLH3 foci along barley (*Hordeum bulgari*) chromosomes during the prophase I. MLH3 participates in the resolution of recombination intermediates, particularly in the formation of crossovers (Colas I. *et al*, 2016). This data was analysed using the gamma distribution method, identifying the MLH3 foci as crossover events. They quantified the strength of interference (k), where a value of k = 1 indicates no interference, >1 indicates positive interference, and <1 indicates negative interference. The performed analysis gave values for k of 1.44 and 1.58 for 2H and 3H, respectively.

The gamma model has not only been used to study interference in barley, but studies with other cereals have been performed. For example, the study performed by Cyrille Saintenac and colleagues in 2009 will be analysed to have a deeper insight into the use of this model.

In this study, they studied the largest hexaploid wheat chromosome, 3B. For this, they used experimental double-haploid (DH) plants. In DH populations, individual gametes are evaluated, thus having independent measures and a higher statistical power. They also generated 10^5 simulated data sets, each one producing N gametes via the Gamma model. The fitting to the model was done for each data set, and the shape parameter k was fitted in male and female meiosis:

- Male meiosis: 1.2, [0-6,1 confidence interval 95%]
- Female meiosis: 3.5 [0,95-12,75 confidence interval 95%]

The difference in the shape parameter was tested using bootstrap analysis and was not statistically significant.

The data obtained with the experimental DH plants was compared with the simulated data adjusted to a gamma distribution in figure 1:

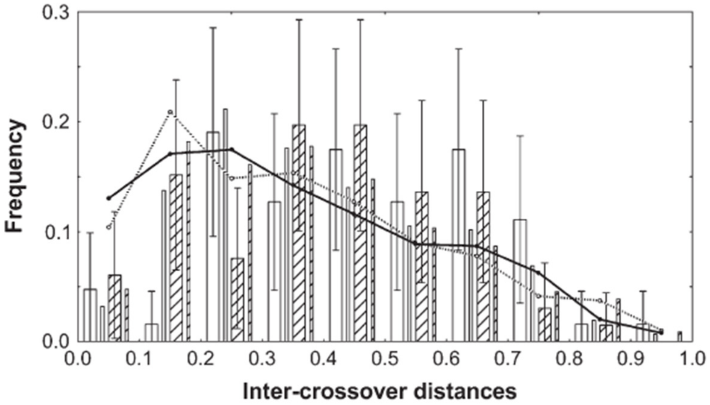


Figure 1. Frequency distribution of distances between two crossovers, referred to as intercrossover distances estimated from recombinant intervals. Open bars, female meiosis; hatched bars, male meiosis; wide bars, experimental data; narrow bars, distribution of simulated gametes generated using the Gamma model. Curves: theoretical expectations under the hypothesis of no crossover interference for male (hatched line, symbols) and female (solid line, symbols) meiosis. Distances are relative to chromosome genetic map length, which is 190 cM in the male map and 194 cM in the female map. (Saintenac, C *et al*. 2009).

These results suggest that the observed DH populations data follows a gamma distribution and that there is existence of crossover interference.

Moreover, an analysis of interference was performed, using the shape parameters presented above. Even if no information about the scale parameter is provided, the shape parameter is the one defining mainly the gamma distribution that fits to the data. The results obtained in the article, based on the coefficient of coincidence, suggested the existence of strong positive interference at distances lower than <10 cM. This interference decreases when distance is increased and becomes weak at distances over 45cM (Saintenac, C *et al*. 2009).

It should be noted that wheat and barley can be crossed to produce *Tritordeum*, a hybrid cereal which has a similar production rate as the wheat, but is more resistant to fungi, floods, and other environmental adversities. The study of crossover events and interference in wheat and barley is relevant to hybrid production. It contributes to the understanding of the genetic basis of these crops and provides tools and insights for breeders aiming to develop hybrids with enhanced traits and performance.

4. SUMMARY AND CONCLUSSIONS

There are several strategies in statistics that can help to build a genetic map and calculate the rate of crossovers and the interference between genes. In this project, three of them were mentioned: chi-square goodness-of-fit, the Count-Location model and the Gamma model. The main aspects are summarized in the following table:

	χ^2 GOODNESS-OF-FIT	**COUNT-LOCATION MODEL**	**GAMMA MODEL**
DISTRIBUTION TYPE	Goodness-of-fit test distribution for observed vs. expected data	Probability distribution for the number of chiasmata.	Probability distribution for crossover interference.
NATURE OF THE MODEL	Tests the fit between observed and expected data.	Describes the distribution of chiasmata on a four-strand bundle.	Describes the distribution of crossover interference along chromosomes.
PARAMETERS	Degrees of freedom (df) in the chi-square distribution.	Parameters may include probabilities (p0, p1, p2…) of 0, 1, 2 chiasmata.	Shape and scale parameters of the gamma distribution.
PROBABILITY FUNCTION	Depends on observed and expected frequencies of crossovers.	Formulation depends on specific assumptions for the CL model.	Depends on the parameters (specially shape) of the gamma distribution.
APPLICATIONS	Hypothesis testing for the fit of observed vs. expected crossovers.	Genetic recombination studies, modelling crossover events.	Genetic studies investigating crossover interference patterns.
CONCEPT OF INTERFERENCE	It can be used to test the presence or absence of interference.	The idea of interference might not be clearly mentioned	The model incorporates the concept of crossover interference.

The use of these statistical tools is highly helpful to analysing inheritance of traits, and can even be used to improve crop production by creating more efficient hybrids and providing a deep knowledge on sexual reproduction.

5. BIBLIOGRAPHY

BarleyWorld. (n.d.). OWB Data. https://barleyworld.org/owb/data. Last visited: 03/12/2023.

Broman, K. W., & Weber, J. L. (2000). Characterization of human crossover interference. American journal of human genetics, 66(6), 1911–1926.

Chi-Square Distribution Table. School of Mathematics and Physics, University of Queensland. https://people.smp.uq.edu.au/YoniNazarathy/stat_models_B_course_spring_07/distributions/chisqtab.pdf. Last visited: 03/12/2023.

Colas, I., Macaulay, M., Higgins, J. D., Phillips, D., Barakate, A., Posch, M., Armstrong, S. J., Franklin, F. C., Halpin, C., Waugh, R., & Ramsay, L. (2016). A spontaneous mutation in MutL-Homolog 3 (HvMLH3) affects synapsis and crossover resolution in the barley desynaptic mutant des10. The New phytologist, 212(3), 693–707.

Saintenac, C., Falque, M., Martin, O. C., Paux, E., Feuillet, C., & Sourdille, P. (2009). Detailed recombination studies along chromosome 3B provide new insights on crossover distribution in wheat (Triticum aestivum L.). Genetics, 181(2), 393–403.

Sapielkin, S., Frenkel, Z., Privman, E., & Korol, A. B. (2022). Statistical analysis and simulation allowing simultaneously positive, negative, and no crossover interference in multilocus recombination data. bioRxiv, 2022-11.