Model of positive selection and fixation of new alleles

Since in *Drosophila* the transcription process takes place even during the post-meiosis period of spermatogenesis, the population genetic model that follows aims to compare the effects of positive selection acting in the haploid (gamete) phase with the corresponding one taking place during the diploid phase, in relation to the important process of positive selection and fixation of new alleles.

The model considers an autosomal locus with alleles **A** and **a** that are expressed in spermatogonia (diploid phase) as well as in gametes (haploid phase). We start by detailing a deterministic model, in which the *Drosophila* population size is very large, matings occur randomly after Hardy-Weinberg ratios and the effects of genetic drift are considered as negligible. The analysis is then followed by the results obtained from an equivalent model, in which the same parameters of the previous model are kept but significant levels of random genetic drift are allowed.

DETERMINISTIC MODEL

(a) Assuming: that $\mathbf{s_1}$ $(\mathbf{0 \le s_1 \le 1})$ and $\mathbf{hs_1}$ $(\mathbf{0 \le h \le 1})$ are the respective coefficients of selection of genotypes **AA** and **Aa**, while **1** is the relative fitness value of genotype **aa;** that $\mathbf{q = q_1}$ and $\mathbf{1-q = 1-q_1}$ are the frequencies of alleles **a** and **A;** and that $\mathbf{q_1'}$ is the frequency value of the **a** allele in the next generation, then it comes out that

$$q_1' = q[1-(1-q)hs_1]/\{1-(1-q)s_1[1-q(1-2h)]\}$$

and

$$\Delta q_1 = q_1' - q = q(1-q)s_1[1-h-q(1-2h)]/ \{1-(1-q)s_1[1-q(1-2h)]\}.$$

In the formulas above, **h** is a dominance measure. When $\mathbf{h = 1}$, it means that the fitness values of **AA**, **Aa** and **aa** genotypes are respectively $\mathbf{W_{AA} = 1-s_1}$, $\mathbf{W_{Aa} = 1-s_1}$ and $\mathbf{W_{aa} = 1}$ and therefore there exists a positive selection mechanism favoring the recessive genotype **aa**. When $\mathbf{h = 0}$, it means that the fitness values of **AA**, **Aa** and **aa** genotypes are respectively $\mathbf{W_{AA} = 1-s_1}$, $\mathbf{W_{Aa} = 1}$ and $\mathbf{W_{aa} = 1}$ and therefore there exists a positive selection mechanism favoring the dominant genotypes **Aa** and **aa**.

(b) Now we let $\mathbf{s_2}$ $\mathbf{(0 \leq s_2 \leq 1)}$ be the coefficient of selection of **A** gametes and **1** the relative fitness of gametes carrying the **a** allele; $\mathbf{q = q_2}$ and $\mathbf{1\text{-}q = 1\text{-}q_2}$ are the population frequencies of **a** and **A** gametes. We let also $\mathbf{q_2'}$ be the frequency value of the **a** allele resulting from gametes that compete among themselves to form the next generation genotypes. Then it comes out that

$$q_2' = q/[1-(1-q)s_2] \;,\; \Delta q_2 = q_2' - q = q(1-q)s_2/[1-(1-q)s_2].$$

(c) Let $\mathbf{\Delta q_2/\Delta q_1}$ be the incremental rate, a pertinent variable for comparing the evolutionary gain of frequency (fixation rate) of the allele **a** under the alternative hypotheses of positive selection acting during the haploid and diploid phases respectively; its value is

$$\Delta q_2/\Delta q_1 = s_2\{1-(1-q)s_1[1-q(1-2h)]\}/\{s_1[1-(1-q)s_2][1-h-q(1-2h)]\}.$$

Instead, if we put $\mathbf{s_2} = \mathbf{s}$ and $\mathbf{s_1} = \mathbf{sx}$, with the obvious restriction $\mathbf{s_1 = sx <= 1}$, we obtain the more suitable expression

$$\Delta q_2/\Delta q_1 = \{1-(1-q)sx[1-q(1-2h)]\}/\{x[1-(1-q)s][1-h-q(1-2h)]\}.$$

(d) Let now **k** ($\mathbf{0 \leq k \leq 1}$) and **1-k** be respectively the contribution proportions of haploid and diploid phases to the transcription process during spermatogenesis. The frequency $\mathbf{Q'}$ of the allele ***a*** as a result of the whole process is obtained by averaging (by **1-k** and **k** respectively) the contributions of diploid and haploid phases in gene frequency $\mathbf{q_1'}$ and $\mathbf{q_2'}$ to the next generation. Since $\mathbf{\Delta Q = Q' - q}$, $\mathbf{\Delta q_1 = q_1' - q}$ and $\mathbf{\Delta q_2 = q_2' - q}$, the immediate result is

$$Q' = kq_2' + (1-k)q_1' = k(\Delta q_2 - \Delta q_1) + \Delta q_1 + q \;,$$
$$\Delta Q = Q' - q = k(\Delta q_2 - \Delta q_1) + \Delta q_1 \;,$$

and

$$\Delta Q/\Delta q_1 = k(\Delta q_2 - \Delta q_1)/\Delta q_1 + 1 = 1 - k(1 - \Delta q_2/\Delta q_1).$$

Since $\mathbf{\Delta Q/\Delta q_1}$ is a linear function of $\mathbf{\Delta q_2/\Delta q_1}$, the behavior of $\mathbf{\Delta Q/\Delta q_1}$ can be straightforwardly (though indirectly) derived from the behavior of $\mathbf{\Delta q_2/\Delta q_1}$. Since the domain of **k** (relative proportion of haploid phase contribution) is $\mathbf{0 < k < 1}$, it comes out that, with no exceptions, $\mathbf{\Delta Q/\Delta q_1 > 1}$ if $\mathbf{\Delta q_2/\Delta q_1 > 1}$ and $\mathbf{\Delta Q/\Delta q_1 < 1}$ if $\mathbf{\Delta q_2/\Delta q_1 < 1}$.

It is essential to keep in mind, in the text below, that the larger the values of the coefficients of selection $\mathbf{s_1 = sx}$ and $\mathbf{s_1h = sxh}$ (of ***AA*** and ***Aa*** genotypes) or $\mathbf{s_2}$ (of gametes **A**), the larger will the relative fitness values of ***aa*** genotypes and **a** gametes be.

The expression derived for the increment rate,

$$\Delta q_2/\Delta q_1 = \{1-(1-q)sx[1-q(1-2h)]\}/\{x[1-(1-q)s][1-h-q(1-2h)]\},$$

can be rewritten in the more convenient form

$$\Delta q_2/\Delta q_1 = 1 + \{1-x[(1-q)(1+hs-h)+qh]\}/\{x[1-(1-q)s][1-h-q(1-2h)]\}.$$

When $\mathbf{0< x} \le 1$, for any combination of values of $\mathbf{0 < s < 1}$, $\mathbf{0 \le h \le 1}$, and $\mathbf{0 < q < 1}$, it comes out that both the numerator and the denominator of the rightmost part of the last equation are positive quantities. Therefore we conclude that if $\mathbf{0 < x \le 1}$, $\mathbf{\Delta q_2/\Delta q_1}$ is always larger than one, irrespectively the value **h** (the dominance factor) can take.

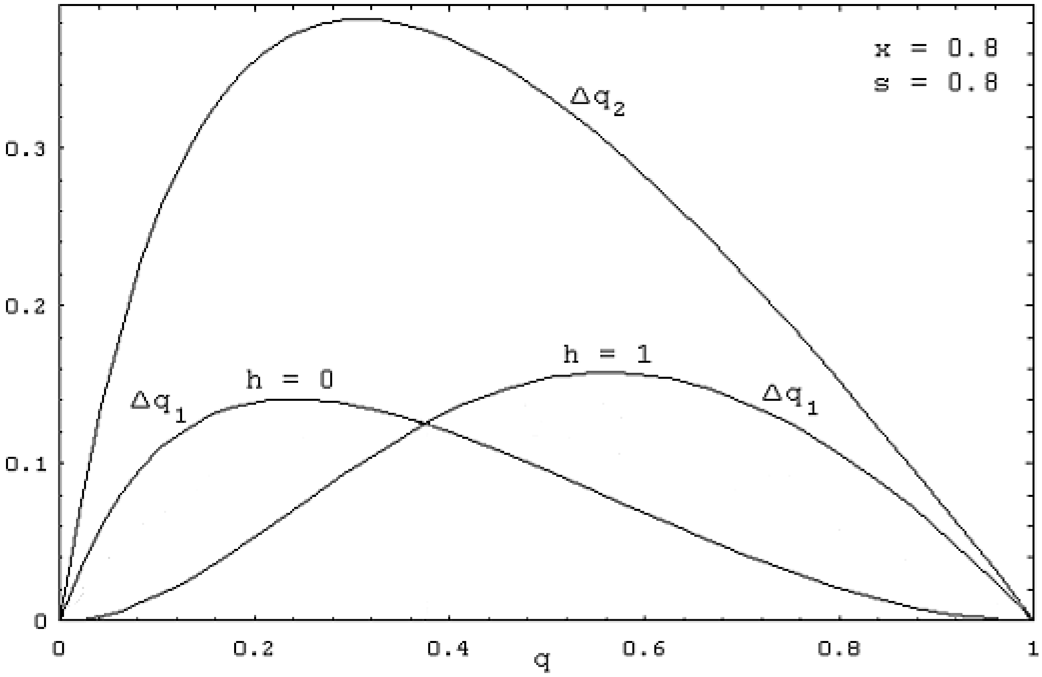

Figure 1 – Comparison between the values of $\mathbf{\Delta q_2}$ and $\mathbf{\Delta q_1}$ in cases **h = 0** and **1**, **x = 0.8** and **s = 0.8**).

The graph of Figure 1 shows unequivocally that for small values of q, the gain in gene frequency $\Delta q_2/\Delta q_1$ (fixation rate) of the ***a*** *allele in the haploid phase is much larger when $h = 1$ (**aa** completely recessive) than when $h = 0$ (dominant case). Extensive computer-assisted numerical analysis showed that this is also true when $h < 1/2$ for all possible combinations of values that q, x and s take inside their corresponding domains $(0, 1)$ and that for small or very small (less than 0.01 or 0.001) frequency values of q the gain in gene frequency $\Delta q_2/\Delta q_1$ when $h = 1$ is about $(1-q)/q$ times larger than when $h = 0$. For example, when $q = 0.001$, the gain in gene frequency of the allele* ***a*** *is about 999 times larger in the case $h = 1$ than in the case $h = 0$.*

When $x > 1$, $\Delta q_2/\Delta q_1$ will be larger or smaller than unity depending on the value (positive or negative) the expression $1-x[(1-q)(1+hs-h)+qh$ from the equation's rightmost part takes, because $x[1-(1-q)s][1-h-q(1-2h)]$ will always be positive, since $x > 1$, $0 \leq h \leq 1$, $0 < q < 1$ and $0 < s < 1/x$ It is important to stress that, unlike the previous case $x \leq 1$, when $x > 1$ it comes out that sx must always be smaller than 1, so that the domain of s is now $0 < s < 1/x$ instead of $0 < s < 1$). Extensive numerical analysis of the formulas above shows that when $h > 1/2$, if $q > [1-x(1-h+hs)]/[x(1-2h+hs)]$, $\Delta q_2/\Delta q_1$ will be smaller than unity and if $q < [1-x(1-h+hs)]/[x(1-2h+hs)]$, $\Delta q_2/\Delta q_1$ will be larger than unity. Otherwise, that is when $h < 1/2$, if $q > [1-x(1-h+hs)]/[x(1-2h+hs)]$, $\Delta q_2/\Delta q_1$ will be larger than unity and if $q < [1-x(1-h+hs)]/[x(1-2h+hs)]$, $\Delta q_2/\Delta q_1$ will be smaller than unity. When $h = 1/2$, the formula of the increment rate reduces to $\Delta q_2/\Delta q_1 = 2[1-sx(1-q)]/[x-sx(1-q)]$ and the numerical analysis of this expression shows that when $x > 1$, for any combination of values of ($0 < q < 1$) and ($0 < s < 1/x$) the increment rate $\Delta q_2/\Delta q_1$ will always be smaller than 1. In fact, replacing x by $1+\delta$, $\delta > 0$, the expression above takes form $\Delta q_2/\Delta q_1 = 2 - 2\delta/\{(1+\delta)[1-s(1-q)]\}$. When s is at its maximum possible value $1/x = 1/(1+\delta)$, the increment rate has value $\Delta q_2/\Delta q_1 = 2 - 2\delta/(\delta+q)$ so that if q is of order of magnitude of δ, $\Delta q_2/\Delta q_1$ will be somewhat smaller than unity; if δ is much larger than q, $\Delta q_2/\Delta q_1$ will be somewhat larger than zero. When s is near its minimum value 0, due to the constraint $s = 1/(1+\delta)$, δ must be very large so that $\Delta q_2/\Delta q_1$ takes a value just a little larger than zero.

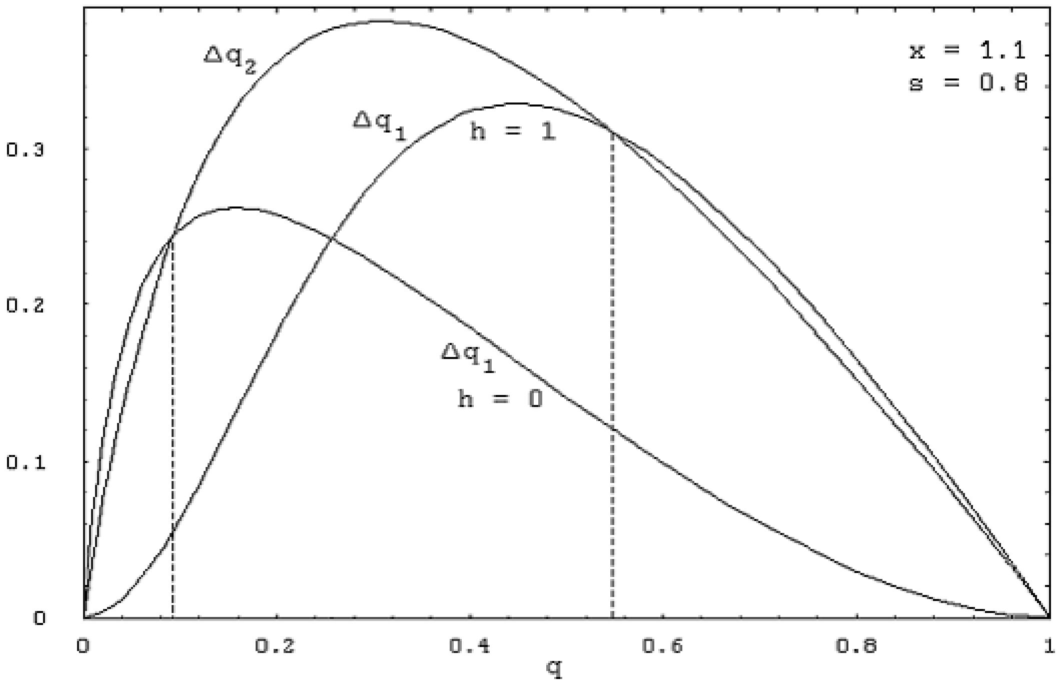

Figure 2 – Comparison between the values of $\Delta \mathbf{q}_2$ and $\Delta \mathbf{q}_1$ in cases **h = 0** and **1, x = 1.1** and **s = 0.8**).

The graph of Figure 2 shows that for small values of q, when ***h = 1*** *(**aa** completely recessive) the gain in gene frequency* $\Delta q_2/\Delta q_1$ *(fixation rate) of the* ***a*** *allele is much larger than in the case* ***h = 0*** *under the system of positive selection during the haploid phase than during the diploid one. This is also valid when other values of* ***h < 1*** *are compared to case* ***h = 1.*** *But in any case for every combination of s and x for some value of q the gain during the diploid phase will be larger than in the haploid phase when* ***x > 1:*** *this takes place, as seen in the above graph, for the prescribed conditions* ***x = 1.1*** *and* ***s = 0.8,*** *when* ***q = (x-1)/x = 1/11 = 0.09091*** *if* ***h = 0*** *and* ***q = (1-sx)/(x-sx) = 6/11 = 0.54545*** *if* ***h = 1.***

Now, if we also take into account the haploid phase relative contribution proportion **k** to the transcription process during spermatogenesis, we finally obtain the (fully) generalized expression

$$\Delta Q/\Delta q_1 = f(q,s,x,h,k) = 1 - k(1 - \Delta q_2/\Delta q_1).$$

Since the domain of $\mathbf{k}$ (relative haploid phase contribution) is $\mathbf{0 < k < 1}$, it comes out straightforwardly that $\mathbf{\Delta Q/\Delta q_1 > 1}$ if $\mathbf{\Delta q_2/\Delta q_1 > 1}$, that is, independently from the relative proportions $\mathbf{k}$ and $\mathbf{1-k}$ of haploid and diploid phase contributions to the transcription process during spermatogenesis, the gain in allele **a** frequency under positive selection during the haploid phase is always larger than the corresponding one during the diploid phase. This situation $\mathbf{\Delta q_2/\Delta q_1 > 1}$ takes place: (1) without any restrictions always when $\mathbf{x \le 1}$, that is when the fitness value of ***a*** gametes is equal or larger than the fitness value of individuals **aa**, independently from the value **h** takes; and (2) with the following restrictions when $\mathbf{x > 1}$ (fitness value of individuals ***aa*** larger than that of ***a*** gametes): $\mathbf{q < [1-x(1-h+hs)]/[x(1-2h+hs)]}$ if $\mathbf{h > 1/2}$, and $\mathbf{q > [1-x(1-h+hs)]/[x(1-2h+hs)]}$ if $\mathbf{h < 1/2}$. If these stringent conditions however do not prevail, the gain in allele ***a*** frequency under positive selection in diploid phase is always larger than the corresponding one in haploid phase.

Table 1 shows the numbers ($\mathbf{n_1}$) and frequency values ($\mathbf{n_1/n}$) of cases $\mathbf{\Delta q_1/\Delta q_2}$ larger than unity that were obtained from $\mathbf{n}$ cases $\mathbf{s_1 > s_2}$ generated by computer-assisted random combinations of $\mathbf{q}$, $\mathbf{s_1 = sx}$, $\mathbf{s_2 = s}$, and $\mathbf{h}$.

Table 1 - Numbers ($\mathbf{n_1}$) and frequencies ($\mathbf{n_1/n}$) of cases $\mathbf{\Delta q_1/\Delta q_2}$ larger than unity obtained from $\mathbf{n}$ cases $\mathbf{s_1 > s_2}$.

n_1	n	n1/n
64	100	0.6400
125	200	0.6250
322	500	0.6440
585	1000	0.5850
3001	5000	0.6002
6142	10000	0.6142
12297	20000	0.6149
30662	50000	0.6132
61489	100000	0.6149
122824	200000	0.6141
306724	500000	0.6134
614072	1000000	0.6141

We conclude therefore that, even in the non-advantageous situation where $x > 1$*, for all possible combinations of* q*,* s*,* x*, and* h *values, in about* ***38.5 %*** *of cases the rate of frequency gain (fixation rate) of the* **a** *allele is larger under the system of positive selection during the haploid phase than during the diploid one (*$\Delta q_2/\Delta q_1 > 1$*). Taking into account that this is exactly what always takes place when* $x \leq 1$*, we have just evidenced the importance of the mechanism of positive selection acting during the haploid phase of spermatogenesis in the process of fixation of new genes.*

ALLOWANCE FOR RANDOM GENETIC DRIFT

In order to take into account the effects of random genetic drift, millions of diploid populations with distinct sizes were computer-simulated. For each diploid population of size N, $2N$ alleles (**a** or **A**) were formed, by means of comparisons with computer-generated random numbers, uniformly distributed and normalized between 0 and 1, and each genotype was formed by a pair of sequentially-generated alleles. The survival of each possible genotype thus created (**aa**, **Aa**, or **AA**) was decided by the comparison of their corresponding relative fitness values ($W_{aa} = 1$, $W_{Aa} = 1-sxh$ e $W_{AA} = 1-sx$) with other computer-generated random numbers. When the genotype did not succeed, a new simulation was performed in order to replace it and thus keep the population size (N) fixed (*soft selection procedure*). The process was thus repeated until 50 populations with N genotypes were formed for each combination of q, s, x and h, with each parameter varying from 0 to 1 with nine fixed intervals of $1/8$ each in the case $x <= 1$. In the alternative case $x > 1$, because of the restriction $sx <= 1$, the procedure was the same for parameters q and h, but used instead both selection parameters $s_1 = sx$ and $s_2 = s$, with x taken indirectly from s_1/s_2. In total, at least $50 \times 9^4 \times 2N = 656{,}100\ N$ computer simulations were performed for each N varying from 5 to $1{,}600$ (*3,280,500 to 1,049,760,000 simulations*) in the case $x <= 1$ and at least $50 \times 9^2\ x\ 10 \times 2N = 81{,}000\ N$ simulations for each N also varying from 5 to $1{,}600$ (*405,000 to 12,960,000*) in the case $x > 1$. From each one out of the 50 genotypic compositions so obtained for each population of size N with a particular combination of the four parameters $\{q, s, h, x\}$, the frequency $q_1' = [2N(aa) + N(Aa)]/2N$ was directly estimated and used to calculate the value of $\Delta q_1 = q_1'-q$, which was then compared to $\Delta q_2 = q(1-q)s/[1-(1-q)s]$ to compute the number of times in which $\Delta q_1 > \Delta q_2$. The value of

$\mathbf{\Delta q_2}$, contrary to what happened to the value of $\mathbf{\Delta q_1}$, was estimated directly from the formula derived in the deterministic model, since the selection in the haploid model obviously results from a practically infinite number of gametes that compete among themselves to form the genotypes of the next *Drosophila* generation.

Table 2 shows, for both cases **x <= 1** and **x > 1**, the results (rounded percentage figures) we obtained for the frequencies of cases in which $\mathbf{\Delta q_1}$ was larger than $\mathbf{\Delta q_2}$ when **50** populations of each size **N** were simulated.

	x <= 1		**x > 1**	
N	**obs. prop.**	**%**	**obs. prop.**	**%**
5	**95850/328050**	**29**	**23494/40500**	**58**
25	**57492/328050**	**18**	**24822/40500**	**61**
50	**46316/328050**	**14**	**23408/40500**	**58**
75	**40358/328050**	**12**	**24482/40500**	**60**
100	**36596/328050**	**11**	**24576/40500**	**61**
200	**30320/328050**	**9**	**25278/40500**	**63**
300	**27052/328050**	**8**	**25654/40500**	**63**
400	**23205/328050**	**7**	**24909/40500**	**62**
600	**22862/328050**	**7**	**25626/40500**	**63**
800	**22285/328050**	**7**	**25542/40500**	**63**
1200	**20908/328050**	**6**	**26288/40500**	**65**
1600	**20392/328050**	**6**	**24830/40500**	**61**

As Figure **3** clearly shows, the percentage figures obtained in the case **x <= 1** correspond with negligible statistical error to the function $\mathbf{y = e^{3.84}/N^{0.30}}$ [**F(1,10) = 1404.30, P = 0.00001, r^2 = 0.993**], which indicates that the percentage value **y** of cases in which the gain (due entirely to random genetic drift) in the diploid phase is larger than in the haploid phase ($\mathbf{\Delta q_1 > \Delta q_2}$) can be obtained directly from this formula. It is not difficult to conclude, however, that even with very large population numbers, on average in around **5%** of cases **x <= 1** the selective gain (fixation rate of the **a** allele) will be larger in the diploid than in the haploid phase. In any case, and for any population number, the number of cases in which the haploid gain predominates is overwhelming in spite of drift.

The percentage figures obtained for the case **x > 1**, on the other hand, indicate that they do not differ significantly (and independently from the population size **N**) from the overall value (Table 1) obtained in the deterministic model (around 61.5%): actually, the average value to all figures shown in Table 2 for the case **x > 1** is exactly 61.5%. We conclude therefore that random genetic drift does not interfere significantly with the dynamics of the deterministic model we described for this specific case.

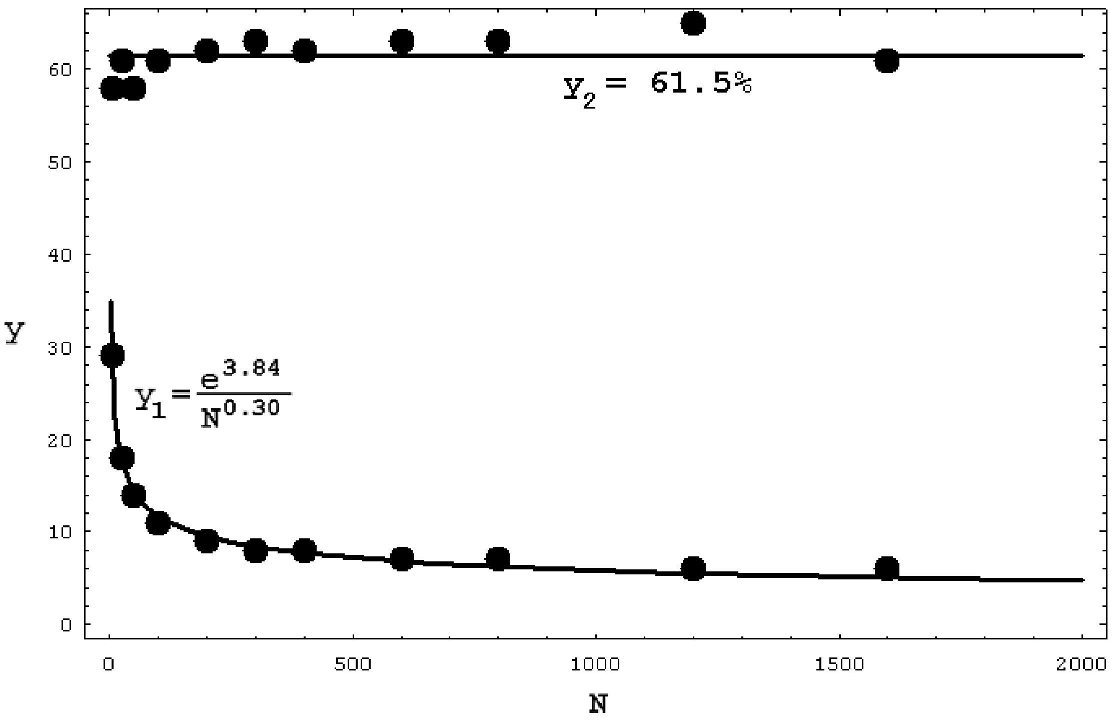

Figure 3 - The black dots represent the percentage of cases in which the selective gain in the fixation process is larger in the diploid than in the haploid phase, due to random genetic drift depending on selection and population size **N**. The upper set and the line $\mathbf{y_2}$ correspond to the case **x > 1** and the lower set and the function $\mathbf{y_1}$ correspond to the alternative case **x <= 1.**

Some parallel results obtained in relation to cases **x = 0.8** and **x = 1.1** (for **s = 0.8** and **h = 0** or **1**) are shown in the set of graphs of Figures 4 and 5, in order to keep the same parameters used in Figures 1 and 2. For this example, 100 (instead of 50) populations of size n (25, 200 or 1000) were generated for 19 reference values of **q** varying from **0.05** to **0.95.** The values of $\mathbf{\Delta q_1}$ were calculated as above and are represented in the graphs of figures 3 and 4 as black dots.

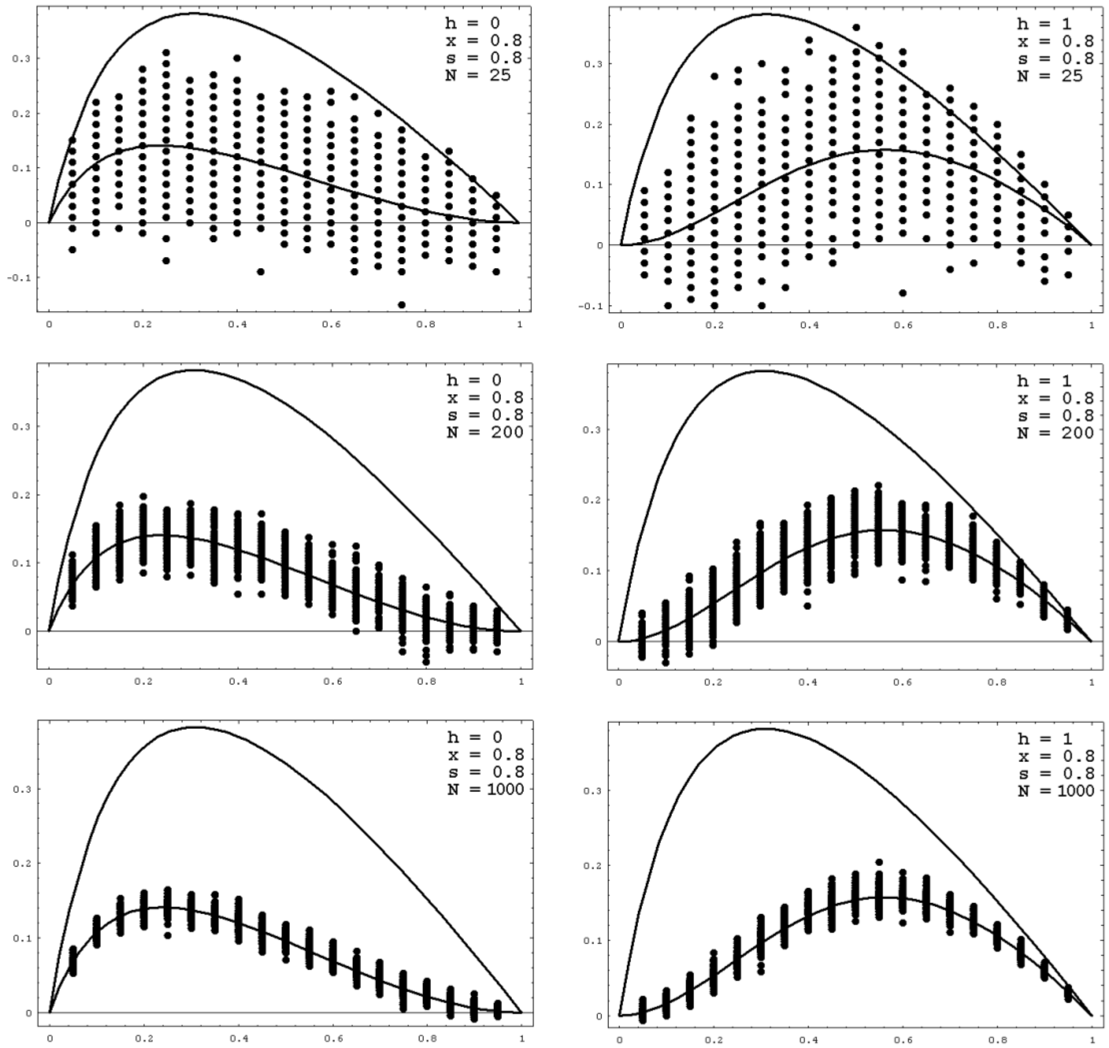

Figure 4 – Results (case **$x < 1$**) obtained from drift/selection simulations for cases **$h = 0$** (right column) and **$h = 1$** (left column) for population sizes of **25** (upper row), **200** (second row) and **1000** (lower row), keeping the parameters as prescribed in Figure **1** (case **$x = 0.8$** and **$s = 0.8$**). The **Δq_1** values obtained from simulated populations are shown as black dots (around the curves representing **Δq_1** in Figure **1**).

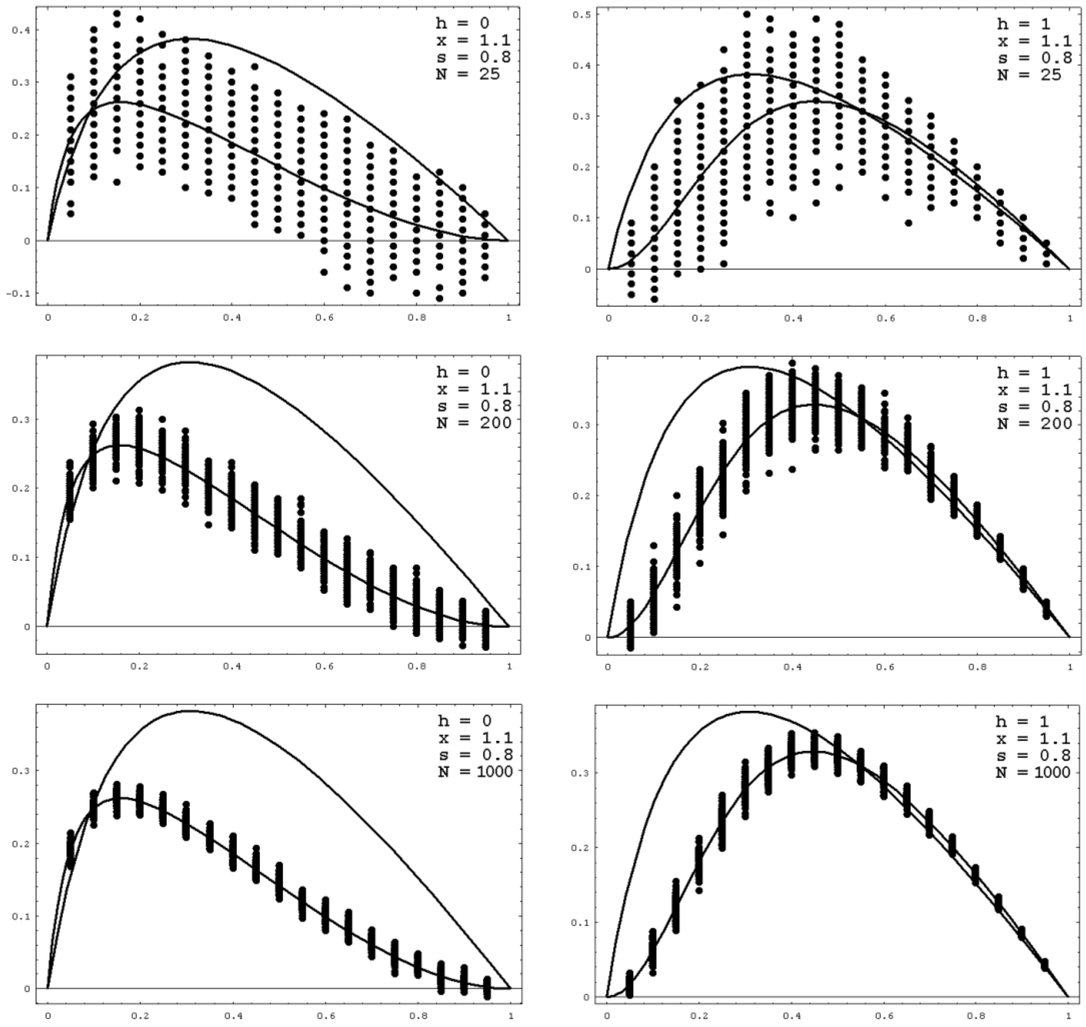

Figure 5 – Results (case **x > 1**) obtained from drift/selection simulations for cases **h = 0** (right column) and **h = 1** (left column) for population sizes of **25** (upper row), **200** (second row) and **1000** (lower row), keeping the parameters as prescribed in Figure **2** (case **x = 1.1** and **s = 0.8**). The Δq_1 values obtained from simulated populations are shown as black dots (around the curves representing Δq_1 in Figure **2**).