Lecture 5

Artificial Selection

Artificial selection, choosing those individuals with the most desirable character values to form the next generation, is arguably one of the most successful enterprises that humans have ever engaged in. Through repeated application of this simple process, our ancestors were able to "domesticate" a wide variety of plant species, transforming them into more stable and higher yielding crops that initiated the rise of societies and, ultimately, modern civilization. The success of artificial selection also provided the critical clue to Charles Darwin in his search for a mechanism underlying natural evolutionary change, ultimately resulting in his theory of evolution by natural selection.

Some examples of the utility of artificial selection:

- Applications in agriculture and forestry - improved yield, production traits in crops and animals
- Creation of model systems of human diseases and disorders (e.g. mouse models of hypertension, obesity, diabetes, behavior)
- Construction of genetically divergent lines for QTL mapping and gene expression (microarray) analysis
- The observed selection response allows for inferences about numbers of loci, effects and frequencies.
- Evolutionary inferences: correlated characters, effects on fitness, long-term response, effect of mutations

Response to Artificial Selection

Selection changes the distribution of a trait. While in theory one could focus on describing the entire change in the distribution, quantitative geneticists typically assume distributions are roughly normal, and hence describing the changes in the mean and variance is sufficient to describe the change of the entire distribution. We will focus largely on changes in the mean.

It is critical to distinguish between the **within-** and **between-generation** changes induced by selection. The within-generation change is the difference in a population before and after an episode of selection, while the between-generation change (the **response to selection**) is the difference between the population distribution before selection and the distribution of the trait in the next generation (measured at the suitable stage). The response to selection depends not only on the strength of within-generation selection, but also on the fraction of offspring trait value that can be predicted from parental value. If the latter is zero, no matter how strong the within-generation selection is, there will be no response to selection.

Truncation Selection

Artificial selection is usually performed by either (i) choosing the uppermost fraction p of the population or (ii) choosing individuals that exceed some threshold value (e.g., disease free). Both of these are examples of **truncation selection**.

The Selection Differential S and Response R

The within-generation change in the mean due to selection is

$$S = \mu_* - \mu_P$$

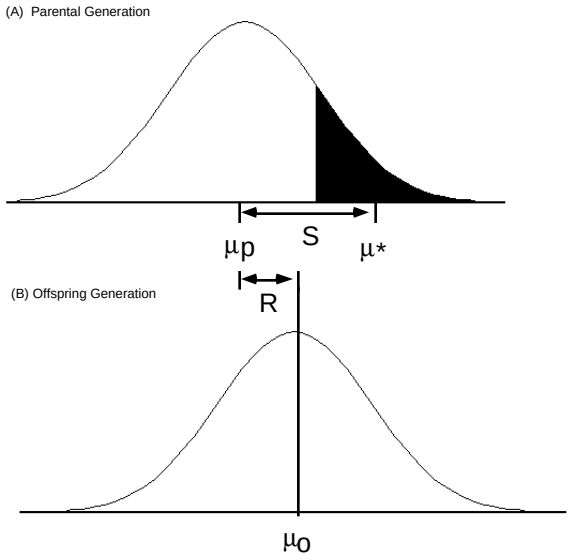
where μ_P is the population mean before selection and μ_* the mean of the parents that reproduce (the population mean after selection). S is called the **selection differential**, or more generally, the **directional selection differential**.

The between-generation change, (the response to selection) R, is the change in means between the population before selection and the population in the next generation,

$$R = \mu_o - \mu_P$$

where μ_o is the character mean in the offspring (measured at the same stage as in their parents).

For example, assume a normal distribution of phenotypic values in Generation 0, of which we select the top 20% ($p = 0.20$) to be parents of the next generation:



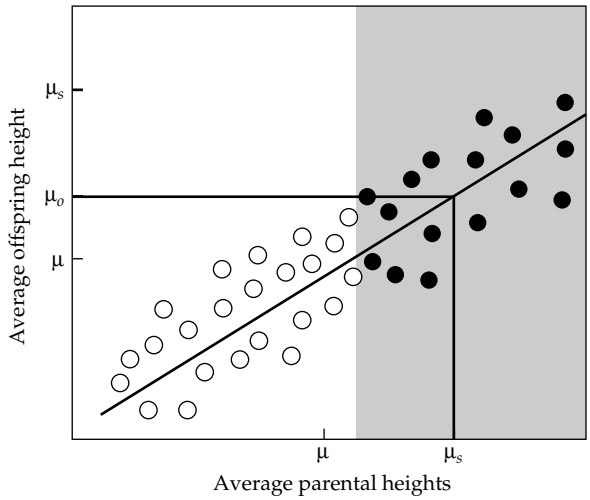
The Breeders' Equation: Translating S into R

The parent-offspring regression allows us to translate the within-generation change S into the between-generation change R. Recall (Lecture 3) that the predicted value $\widehat{y}$ given we know x is

$$\widehat{y} = \mu_y + b_{y\,|\,x}(\,x - \mu_x\,)$$

Here we are trying to predict the offspring value y_O given $x = (P_f + P_m)/2$, the midparent value. Hence, $b_{y\,|\,x} = b_{O\,|\,MP} = h^2$ is the slope of the midparent-offspring regression, while $\mu_y = \mu_x = \mu_P$, the mean trait value in the population, giving

$$y_O = \mu_P + h^2\left(\frac{P_f + P_m}{2} - \mu_P\right)$$



This regression holds for each midparent-offspring pair. Averaging over all parents, the average difference between the selected parents and the (before selection) population mean is

$$E[\,(P_f + P_m)/2 - \mu_P] = \mu_* - \mu_P = S$$

Likewise, the average value over all the offspring of these selected parents is $E[\,y_O\,] = \mu_O$. Thus, averaging over all the midparents gives

$$\mu_O = \mu_P + h^2\, S$$

since $R = \mu_O - \mu_P$, this gives

$$R = h^2 S \tag{1}$$

This relationship is often called the **breeders' equation**, and shows that the heritability of a character is the link between the within-generation change S and the between-generation response R. If $h^2 \simeq 0$, then $R \simeq 0$ no matter how strong the amount of selection is applied

If selection is practiced among a population of clones (so that offspring are genetically identical to their parents), h^2 is replaced by $H^2 = V_G/V_P$, the broad-sense heritability, as this is the slope of the parent-offspring regression when offspring are asexual clones of their parents. Such selection among clones is common in plant breeding, where selection occurs among a collection of fully-inbred lines.

In some situations, males and females are subjected to different amounts of selection. In this case, the selection differential is simply the average differential of both sexes,

$$S = \frac{S_f + S_m}{2}$$

For example, in many crop plants selection for traits occurs *after pollination*, e.g., using seeds from the tallest plants to form the next generation. This has the result of no selection on the pollen (male) parent ($S_m = 0$), giving the total amount of selection as $S = S_f/2$, half the amount of selection on the females.

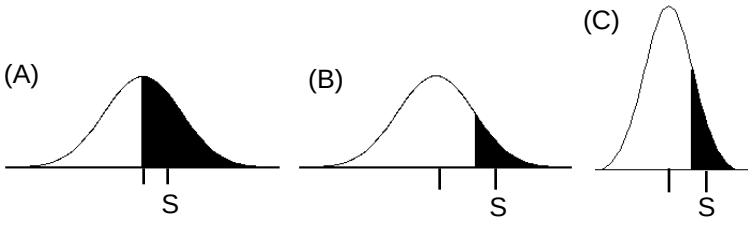
While the breeders' equation holds for a single generation of selection from an unselected base population, its validity in predicting response over several generations depends on:

- The reliability of the h^2 estimate
- Absence of environmental change between generations
- The absence of genetic change between the generation in which h^2 was estimated and the generation in which selection is applied.

The later point is critical, as strictly speaking, the prediction equation is true for one generation only, since selection changes gene frequencies and thus h^2 (through changes in the genetic variances). In practice, the prediction equation is generally valid over several generations.

Computing S Under Truncation Selection

The selection differential, S, will not be exactly known until the selection has already been made among the parental generation. Hence, as it stands the breeders' equation has a limited value in trying to plan future selection programs/experiments. Fortunately, if we practice truncation selection, we can determine S in advanced by our choice of the fraction saved p. This follows since, for truncation selection, we can express S as a function of (i) p, the proportion selected and (ii) the phenotypic variance, $V_P = \sigma_P^2$. As we decrease the fraction saved p, we increase the selection differential S. Likewise, for the same fraction p saved, S increases as the phenotypic variance increases.



(a) 50% selected, $\sigma_P^2 = 4, S = 1.6$

(b) 20% selected, $\sigma_P^2 = 4, S = 2.8$

(c) 20% selected, $\sigma_P^2 = 1, S = 1.4$

Since the selection differential is a function of the phenotypic variance V_P, two populations can show the same S value, yet selection is much stronger (i.e., p is much smaller) in one of the populations because it has a much smaller variance. To compare the amount of selection independent of the phenotypic variance, the **selection intensity** i is used, where

$$i = \frac{S}{\sqrt{V_P}} = \frac{S}{\sigma_p}$$

i is also called the **standardized selection intensity**.

Assuming the trait is normally distributed, i is solely a function of the proportion selected (p), with i increasing as p decreases. Values of i as a function of p have been tabulated (F&M Appendix Tables A and B, pp. 379-380). If the number of selected individuals is large, then an approximation for i is given by

$$i \simeq 0.8 + 0.41 \ln\left(\frac{1}{p} - 1\right)$$

This approximation is generally quite good for $0.004 \leq p \leq 0.75$.

The **selection intensity** version of the breeders' equation follows since

$$R = h^2 S = h^2 \frac{S}{\sigma_p} \sigma_p = i\, h^2\, \sigma_p$$

which (recalling $h^2 = \sigma_A^2/\sigma_P^2$) this can also be rewritten as

$$R = i \frac{\sigma_A^2}{\sigma_p}$$

If the proportion of males and females selected differ (i_m and i_f),

$$\overline{i} = \frac{i_m + i_f}{2}$$

Measurement of Response to Selection

1. Variability of generation means

Selection over the course of several generations typically proceeds more or less erratically.

The variation in generation means is caused by

(i) random genetic drift

(ii) sampling error in estimating generation means

(iii) differences in the selection differential

(iv) changes in the environment over time.

To control for environmental fluctuations, it is necessary to maintain a contemporaneous population of the same size that is either unselected (a **control population**) or selected for the trait in the opposite direction (**divergent selection**). Assuming no GxE, the environmental changes affect the selected and control (divergent) lines equally. To estimate the response, the deviation of the selected and control lines are computed each generation.

To control for effects of drift, experiments designed to measure response must be replicated.

2. Realized heritability

The most useful empirical measurement of the effectiveness of selection is the **realized heritiability**, h_r^2, which follows from a simple rearrangement of the breeders' equation:

$$h_r^2 = \frac{R}{S} = \frac{\text{Observed Response}}{\text{Selection Differential}}$$

Realized heritability h_r^2 allows comparison of different experiments and is the best predictor of what would happen if one repeated the experiment.

If our selection experiment spans multiple generations, the realized heritability is estimated using the **cumulative selection response**

$$R_C(t) = \sum_{i=1}^{t} R(i)$$

which is the total response after t generations of selection and the **cumulative selection differential**

$$S_C(t) = \sum_{i=1}^{t} S(i)$$

which is the total selection over the t generations. If our experiment runs a total of T generations, the simplest estimator of the realized heritability is just the ratio of total selection response to total selection differential,

$$\widehat{h}_r^2 = \frac{R_C(T)}{S_C(T)}$$

This is the **ratio estimator** of the realized heritability.

A much more widely used approach is the **regression estimator**, the slope of the regression of cumulative response on cumulative differential,

$$R_C(t) = h_r^2 \, S_C(t) + e_t$$

here e_t is the residual for generation t, the difference between the predicted [$h_r^2 \, S_C(t)$] and observed [$R_C(t)$] response. Two technical comments about this regression. First, it is forced through the origin, as if there is no selection differential, there is no response. In such cases of regressions through the origin, the estimate of the slope becomes

$$\widehat{b} = \frac{\sum_t S_C(t) R_C(t)}{\sum_t S_C(t)^2}$$

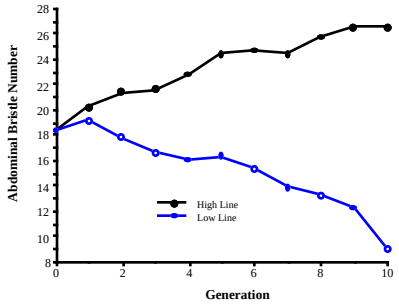
Second, the residuals tend to be correlated (this effect is generated by drift), so that if one wishes to estimate the standard error on the slope (i.e., the SE for the realized heritability), a weighted least-squares approach must be used. Unfortunately, ordinary (i.e., unweighted) least squares is generally used, resulting in a serve *underestimation* of the standard error, giving the false impression of higher precision than is actually present.

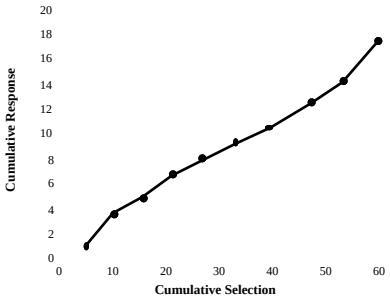
Example: Computing Realized h^2 from Selection Response

- The experiment: Select for increased and decreased abdominal bristle number in *Drosophila melanogaster*. The base population consisted of 62 isofemale lines from Raleigh Farmers Market population, crossed in a round robin design so that each family contributes equally.
- Select 25 males, 25 females with highest bristle number from 100 scored of each sex ($p = 25\%$, $i = 1.26$) in high selection line, and the same in the low selection line. Selection was continued 25 generations, the first 10 of which are considered here.
- Each generation, record the mean of all 200 progeny (μ_o) and the mean of the selected group (μ_*). From these data, compute $R(t) = \mu(t+1) - \mu(t)$ and $S(t) = \mu_*(t) - \mu_P(t)$, and cumulate the total response and selection each generation.

The data are as follows:

	High Lines				Low Line				Divergence	
t	μ_p	μ_o	$R_c(t)$	$S_c(t)$	μ_p	μ_o	$R_c(t)$	$S_c(t)$	$R_c(t)$	$S_c(t)$
0	18.3	20.8	1.8	2.5	18.3	15.8	0.8	-2.5	1	5
1	20.1	22.9	3	5.3	19.1	16.7	-0.6	-4.9	3.6	10.2
2	21.3	23.9	3.2	7.9	17.7	14.8	-1.7	-7.8	4.9	15.7
3	21.5	24.3	4.4	10.7	16.6	14	-2.3	-10.4	6.7	21.1
4	22.7	25.8	6	13.8	16	13.4	-2	-13	8	26.8
5	24.3	27.4	6.3	16.9	16.3	13.3	-3	-16	9.3	32.9
6	24.6	27.8	6	20.1	15.3	12.1	-4.5	-19.2	10.5	39.3
7	24.3	29.1	7.4	24.9	13.8	10.8	-5.1	-22.2	12.5	47.1
8	25.7	28.6	8.1	27.8	13.2	10.1	-6.1	-25.3	14.2	53.1
9	26.4	29.2	8.1	30.6	12.2	8.5	-9.3	-29	17.4	59.6
10	26.4				9					





There are three comparisons for estimating the realized heritability:

Comparison	Regression estimator	Ratio estimator
High line (H)	h_r^2 = 0.223	h_r^2 = 0.265
Low line (L)	h_r^2 = 0.322	h_r^2 = 0.321
Divergence (H-L)	h_r^2 = 0.270	h_r^2 = 0.292

Gene Frequency Changes Under Selection

How quickly does selection change the frequency of alleles at loci contributing to a trait under selection? We start by reviewing a few results from population genetics. Consider a diallelic locus, with alleles A_1 and A_2, whose genotypes have the following relative fitnesses:

Genotype	A_1A_1	A_1A_2	A_2A_2
Fitness	1	$1+s$	$1+2s$

This is an example of **additive fitness**. With these fitnesses, for every offspring left by an individual with an A_1A_1 genotype, $1+2s$ offspring are left (on average) by individuals with an A_2A_2 genotype. If q represents the frequency of allele A_2 before selection, then the change in the frequency of q after selection is given by

$$\Delta q = \frac{sq(1-q)}{1+2sq} \simeq sq(1-q) \quad \text{when} \quad |\,2sq\,| << 1$$

Thus, under these fitnesses, the change in the frequency of the favorable allele is proportional to s. In finite populations, genetic drift can overpower the effects of selection. In particular, when

$$4N_e\,|\,s\,| << 1$$

the fate of an allele is largely determined by gene drift, rather than selection. In such cases, favorable alleles can easily be lost by drift.

Now consider a locus contributing to a character under selection. Supposes the genotypes at this locus make the following contribution to the character:

Genotype	A_1A_1	A_1A_2	A_2A_2
Contribution	0	a	$2a$

For a trait with phenotypic variation σ_z^2 under selection intensity i, this induces additive fitnesses on these genotypes, with

$$s = \frac{a}{\sigma_z}\, i$$

Hence, the change in allele frequency depends on both the strength of selection i and the relative contribution a/σ_z of the character to the overall trait value. As expected, loci with larger contributions are under stronger selection than loci with minor contributions and hence have faster allele frequency changes. Further note that if

$$4N_e|\, s\,| = \frac{4N_e|\, a\, i\,|}{\sigma_z} << 1$$

then the effect of selection on this locus is weaker than the effects of drift. Thus, many favorable QTL alleles can be lost by drift if either their effects (a/σ_z), the strength of selection on the character (i), or the effective population size (N_e) are sufficiently small.

More generally, if the locus shows dominance towards the character, the fitnesses become

Genotype	A_1A_1	A_1A_2	A_2A_2
Contribution	0	$a(1+k)$	$2a$
Induced fitness	1	$1+s(1+h)$	$1+2s$

where for the induced fitnesses $s = ai/\sigma_z$(as above) and $h = k$.

Changes in the Variance

Selection has two routes by which to change the genetic variances, and hence the heritability and selection response. First, it can change the frequencies of individual alleles. When the contribution to a trait from any locus is very small, these selection-induced changes in allele frequencies over a few generations are also very small. However, selection also creates correlations between alleles at different loci (**linkage disequilibrium**), and this can result in an immediate change in the variance.

Consider the within-generation change in the variance, $\delta\sigma_z^2 = \sigma_{z^*}^2 - \sigma_z^2$. Using regression arguments similar to those leading to the breeders' equation, the expected response in the variance to a single generation of selection is

$$d = \sigma_O^2 - \sigma_P^2 = \frac{h^4}{2}\,\delta\sigma_z^2 \tag{12.33}$$

where σ_O^2 is the variance in the offspring and σ_P^2 the variance in the unselected population. Equation 12.33 is the variance response analog to the response in mean (the breeders' equation), with $h^4/2$ replacing h^2 and $\delta\sigma_z^2$ replacing S. In many situations (such as truncation selection), we can write

$$\sigma_{z^*}^2 = (1-k)\sigma_z^2$$

so that the result of selection is a proportional change in the variance.

It turns out that all the change in the variance is due to a change in the additive genetic variance, so that if V_a denotes the additive variance before selection, then after one generation of selection

$$V_A(1) = V_a + d, \quad V_P(1) = V_A(1) + V_D + V_E = V_P + d$$

where V_P is the phenotypic variance in the base (pre-selection) population. The heritability thus becomes

$$h^2(1) = \frac{V_A(1)}{V_P(1)} = \frac{V_a + d}{V_P + d}$$

Truncation selection reduces the variance ($\delta\sigma_z < 0$), which results in reduced additive genetic variance and heritability in the next generation, slowing response. This reduction in variance due to selection creating linkage disequilibrium is referred to as the **Bulmer effect**, after Michael Bulmer's pioneering work on this subject in the 1970's.

One subtle feature of changes in the variance is that recombination breaks down the selection-induced correlations, so that in the absence of selection, $d(t+1) = d(t)/2$. Hence, one must iterate to obtain the value of the variance in generation t. Starting with an unselected base population, $d(0) = 0$, we obtain the value for $d(t+1)$ by iterating

$$\begin{aligned} d(t+1) &= \frac{d(t)}{2} + \frac{h^4(t)}{2}\,\delta\sigma_{z(t)} \\ &= \frac{d(t)}{2} - k\frac{h^4(t)}{2}\,\sigma_z^2(t) \end{aligned}$$

The first term ($d/2$) is the decay in linkage disequilibrium from recombination while the second term is the amount of new disequilibrium created by selection. Note from above that

$$\sigma_z^2(t) = \sigma_z^2(0) + d(t), \qquad \text{and} \qquad h^2(t) = \frac{V_A(t)}{V_P(t)} = \frac{V_a + d(t)}{V_P + d(t)}$$

While all this looks rather complicated at first glance, its really a very straight forward series of substitutions. The net result for directional selection is that most of the reduction in variance occurs over the first few generations, which rapidly approaches an equilibrium value (the equilibrium reduction in the additive variance). However, under disruptive selection (selection to increase the variance, for example by selecting both the largest and smallest parents), the variance may continue to increase substantially over many generations before settling on its equilibrium value.

Artificial Selection Problems

1. Taking the selection differential as the difference between the means of selected parents and the mean before selection makes the assumption that each selected parent contributes equally to the next generation. Biases introduced by differential fertility can be removed by using **effective selection differentials**, S_e,

$$S_e = \frac{1}{n_p}\sum_{i=1}^{n_p}\left(\frac{n_i}{\overline{n}}\right)(z_i - \mu_z) = \left(\frac{1}{n_p}\sum_{i=1}^{n_p}\left(\frac{n_i}{\overline{n}}\right) z_i\right) - \mu_z$$

where z_i and n_i are the phenotypic value and total number of offspring of the ith parent, n_p the number of parents selected to reproduce, $\overline{n}$ the average number of offspring for selected parents, and μ_z is the mean before selection. If all selected parents have the same number of offspring ($n_i = \overline{n}$ for all i), then S_e reduces to S. However, if there is variation in n_i among selected parents, S_e can be considerably different from S. This corrected differential is occasionally referred to as the **realized selection differential**.

Suppose 5 parents are selected, with the following trait values and offspring number:

Parent	phenotypic value	number of offspring
1	45	1
2	40	2
3	35	3
4	33	5
5	32	5

If the mean before selection is 30, compute the S and S_e. If $h^2 = 0.3$, what is the expected response that would be estimated under the two differentials?

2. Consider a population not currently under selection, with $\sigma_z^2 = 100$ and $h^2 = 0.5$ and $d(0) = 0$ (no disequilibrium). Consider two types of selection (i) stabilizing where $\sigma_{z*}^2 = 0.5\sigma_z^2$ (i.e., $k = 1/2$) and (ii) disruptive selection $\sigma_{z*}^2 = 1.5\sigma_z^2$ ($k = -1/2$). For both types of selection compute $d(1)$ and $d(2)$, $\sigma_A^2(1)$ and $\sigma_A^2(2)$, $\sigma_z^2(1)$ and $\sigma_z^2(2)$, and $h^2(1)$ and $h^2(2)$.

Solutions to Artificial Selection Problems

1. Here $\mu_* = 37$, giving $S = 7$, while $\overline{n} = 3.2$ and

i	z_i	n_i	$n_i/\overline{n}$	$z_i \cdot n_i/\overline{n}$
1	45	1	0.3125	14.06
2	40	2	0.6250	25.00
3	35	3	0.9375	32.81
4	33	5	1.563	51.56
5	32	5	1.563	50.0

$$\frac{1}{n_p}\sum_{i=1}^{n_p}\left(\frac{n_i}{\overline{n}}\right) z_i = 34.69$$

Giving $S_e = 4.69$. Assuming $h^2 = 0.3$, using the uncorrected S gives a response of $R = 0.3 \cdot 7 = 2.1$, while the true expected response if $R = 0.3 \cdot 4.69 = 1.4$

2. Here $\sigma_a^2 = h^2\sigma_z^2 = 50$, and $d(0) = 0$

$$d(1) = d(0) - k(h^4/2)\sigma_z^2(0) = \begin{cases} 0 - 0.5 * 0.125 * 100 = -6.25 & \text{for stabilizing, } k = 0.5 \\ 0 + 0.5 * 0.125 * 100 = 6.25 & \text{for disruptive, } k = -0.5 \end{cases}$$

$$\sigma_A^2(1) = \sigma_a^2 + d(1) = \begin{cases} 43.75 & \text{for stabilizing} \\ 56.25 & \text{for disruptive} \end{cases}, \quad \sigma_z^2(1) = \sigma_z^2 + d(1) = \begin{cases} 93.75 & \text{for stabilizing} \\ 106.25 & \text{for disruptive} \end{cases}$$

$$h^2(1) = \sigma_A^2(1)/\sigma_z^2(1) = \begin{cases} 0.467 & \text{for stabilizing} \\ 0.529 & \text{for disruptive} \end{cases}$$

$$d(2) = d(1)/2 - k(h^4(1)/2)\sigma_z^2(1) = \begin{cases} -6.25/2 - 0.5(0.46^2/2) * 93.57 = -8.08 & \text{for stabilizing} \\ 6.25/2 + 0.5(0.53^2/2) * 106.25 = 10.59 & \text{for disruptive} \end{cases}$$

$$\sigma_A^2(2) = \sigma_a^2 + d(2) = \begin{cases} 41.92 & \text{for stabilizing} \\ 60.59 & \text{for disruptive} \end{cases}, \quad \sigma_z^2(2) = \sigma_z^2 + d(2) = \begin{cases} 91.92 & \text{for stabilizing} \\ 110.6 & \text{for disruptive} \end{cases}$$

$$h^2(2) = \sigma_A^2(2)/\sigma_z^2(2) = \begin{cases} 0.456 & \text{for stabilizing} \\ 0.548 & \text{for disruptive} \end{cases}$$