Optimizing breeding schemes //

Manual

Selection intensity

- This manual describes selection intensity, how to calculate it, and its importance in determining response to selection.
- Published on 26/10/2020

excellenceinbreeding.org/toolbox/tools/eib-breeding-scheme-optimization-manuals

Excellence in Breeding PLATFORM

ExcellenceinBreeding.org

Selection intensity

Authors

Ian Mackay / i.j.mackay@gmail.com

IMplant Consultancy Ltd., Chelmsford, CM2 6HA, UK

Editors

Giovanny E. Covarrubias-Pazaran / g.covarrubias@cgiar.org

Breeding Optimization Lead, CGIAR Excellence in Breeding Platform (EiB)

Sam Storr / EiB

CGIAR Excellence in Breeding Platform

Contents

Introduction

Selection intensity is one of four terms in the full form of the breeders' equation:

$$\Delta R = h^2 S/t = ih^2\sigma_p/t = ih\sigma_g/t$$

Where:

ΔR	is the rate of change in response to selection
S	is the selection differential
σ_p	is the phenotypic standard deviation
σ_g	is the genetic standard deviation
t	is the cycle time
i	is the selection intensity

As selection intensity increases, response to selection increases. To understand how to optimize breeding programs it is important to understand some properties of the term. In particular:

1. How to calculate selection intensity.
2. The non-linear relationship of selection intensity with the proportion selected.

This manual describes selection intensity in detail, with calculation methods and examples, in addition to the consequences of varying it within breeding programs.

1. Standardizing variables.

It is difficult to compare variables measured on different scales. When the same property is being measured by variables using different scales, then variables can be easily converted to the same scale, for example, temperatures measured in centigrade and Fahrenheit. To compare variables for height and weight, for example, comparison is not so straightforward. This is usually the case in breeding applications.

A common statistical solution is to standardize all variables to a mean of zero and a variance of one. This is usually done using the estimates of means and variances from the samples under study, but this need not be the case (and may produce misleading results with small sample sizes).

The standardized variable is $\frac{x-\mu}{\sigma}$ where x is an observation, μ is the mean of the distribution and σ is the standard deviation. A standardized distribution need not be a normal distribution, but the mean will be 0 and the variance 1 after standardizing.

For example, a set of ten cereal lines have the following grain weights (t/ha) and heights (cm):

Table 1. Example dataset of ten cereal lines, and standardized units for grain weight and height.

	Raw		Standardized	
ID	**weight**	**height**	**weight**	**height**
1	10.9	49.8	1.25	-0.94
2	10.3	50.1	0.18	-0.93
3	9.6	48.5	-1.08	-0.99
4	9.9	50.9	-0.54	-0.9
5	10	48.5	-0.36	-0.99
6	11.2	99.9	1.79	0.97
7	9.8	99.1	-0.72	0.94
8	10.7	99.1	0.90	0.94
9	9.6	99.7	-1.08	0.96
10	10	99.2	-0.36	0.94
mean	10.2	74.5	0	0
variance	0.31	690.56	1	1
sd	0.56	26.28	1	1

The standardized weight of the first variety is (10.9 – 10.2) / 0.56 or 1.25.

Provided the estimates of the means and variances used in standardizing are recorded, the original variables can be recovered by multiplying by the standard deviation and adding the mean. For example, the original weight for the first variety is 1.25*0.56 + 10.2 = 10.9.

Here, we have standardized using estimates of the mean and variance from the data. Sometimes, better estimates are available, especially if the sample size is small. In that case, the mean and variance after standardizing will no longer be 0 and 1 respectively.

Standardization in this manner is very common in statistics. Relevant examples for breeding include:

i. PCO and PCA analyses: traits are usually standardized before analysis. This is equivalent to working on the correlations between traits rather than the covariances (important note: check what your software uses as a default).
ii. Kinship calculation: marker scores are commonly standardized first.

2. Obtaining selection intensity

The selection differential (S) is the difference between the mean of the selected group and the mean of the population. Selection intensity (i) is therefore the mean of the deviations from the population mean, measured in units of the phenotypic standard deviation of the population.

$h^2S / t = h^2 i\sigma_p / t$

$S = i\sigma_p$

$i = S / \sigma_p$

$i = [\Sigma(y_i - \hat{y}) / n] / \sigma_p$

i = the standardized mean of the selected group

Where:

y_i is the i^{th} *selected* observation

$\bar{y}$ is the *population* mean

n is the number of *selected* observations.

σ_p^2 is the population variance (σ_p is the population standard deviation)

Note that the selected group is standardized by the mean and standard deviation of the whole population and not just by members of the select group.

3. Estimating selection intensity from the properties of the normal distribution

Mathematically, selection intensity (i) is expressed as:

$i = \Phi(x)/p$

Where:

p is the proportion selected, assuming truncation selection.

x is the mean deviation of the selected group $\Sigma(y_i - \hat{y}) / n$

$\Phi(x)$ is the probability density function (i.e. the value on the y axis) of the standard normal distribution for x where:

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

Although this seems complicated, it can be computed very simply:

In R: i <- dnorm(qnorm(1-p))/p

In Excel: =NORMDIST(NORMSINV(1-p) ,0,1,0)/p

In R, p is the variable containing the proportion selected, whereas in Excel it refers to the cell containing the proportion selected.

For example in R:

```
> i <- dnorm(qnorm(1-0.1))/0.1
> i
[1] 1.754983
```

In Excel:

```
=NORMDIST(NORMSINV(1-0.1),0,1,0)/0.1
1.754983319
```

For this reason, it is no longer necessary to use selection intensity reference tables supplied in older textbooks (e.g. Falconer and Mackay, 1996).

4. Relationship between selection intensity and the proportion selected

Figure 1 demonstrates the relationship between selection intensity and the proportion selected. The proportion selected is on a -log10(p) scale where 1 represents the best 10%, 2 the best 1% and 6 the best 0.0001 % (1 in a million). All other things being equal, the breeders' equation shows that response to selection is directly related to selection intensity, but that the relationship with the proportion selected is not linear.

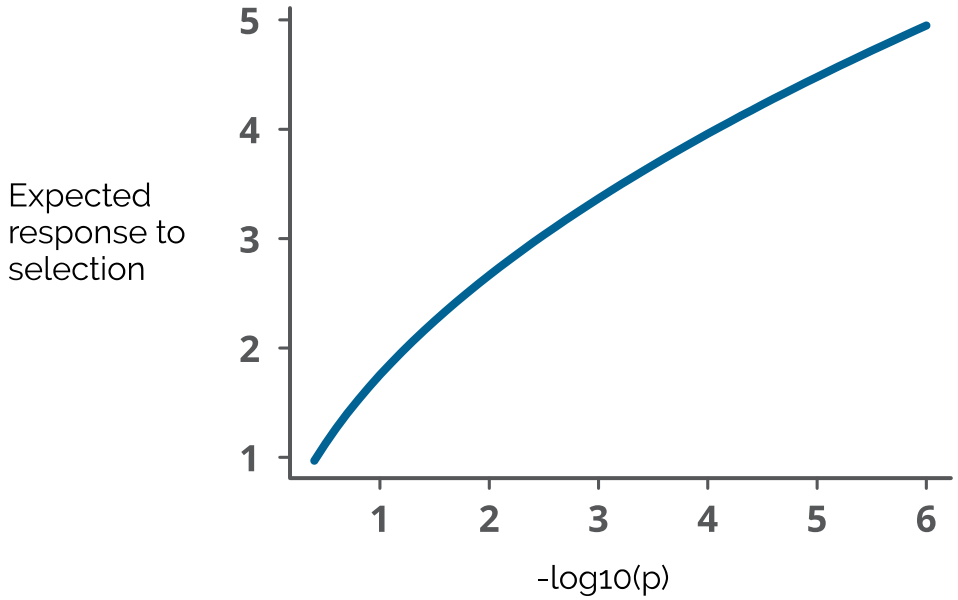

Figure 1. Relationship between the selection intensity and the proportion selected.

Some representative values are provided in **Table 2**. The ratio between intensity is shown for pairs of values representing a doubling of the proportion selected. As the percent selected reduces, the relative gain in intensity, and therefore in response to selection, is consistently reduced.

The last pair of values show a gain of only 3% from decreasing the proportion selected from two in a million to one in a million. A two-fold decrease in proportion selected requires a two-fold increase in the number of candidates for selection to maintain population sizes after selection, which is likely to be expensive. This suggests that investment might be better spent elsewhere. For example, if the cycle time were halved, all other things being equal, the rate of genetic gain is doubled.

Table 2. Relationship between selection intensity and percent selected

% selected	i	Ratio
20	1.4	
10	1.755	1.25
2	2.421	
1	2.665	1.1
0.2	3.17	
0.1	3.367	1.06
0.0002	4.811	
0.0001	4.948	1.03

Using the data in **Table 2**, we can crudely compare a breeding program selecting 1% in a two year cycle (i=2.665 per cycle or 1.333 per year) with one selecting 20% in a two year (i = 1.4 per year and per cycle). The 20% scheme is more effective while the number of lines or individuals tested per year is ten times smaller (10x not 20x because lines must be created and tested twice as often).

Of course, this crude comparison is a gross simplification, which is why computer simulation is used for more sophisticated methods of optimizing breeding programs. However, it does illustrate the diminishing returns from selecting ever harder. This most important lesson of quantitative genetics for breeders is not as well-known as it should be. Rather than increasing the scale of breeding indefinitely, it is more important to focus on doing it better: smarter not bigger.

5. Dependency of selection intensity on population size

There is a dependency of selection intensity on population size. This arises from normal distribution theory in which proportions vary continuously, whereas in reality they vary in steps of 1/n where n is the population size before selection. An approximate correction for this (Bulmer 1980) is to replace the proportion selected in the formula for i given before with:

$$p^* = (k+1/2)/(n+k/2n)$$

Where k and n are the numbers before and after selection, respectively.

Table 3 gives some examples.

Table 3. Relationship between selection intensity and percent selected.

% selected	k	n	p*	i
0.1	1	1,000	0.00150	3.253
	10	10,000	0.00105	3.354
	100	100,000	0.00101	3.364
	∞	∞	0.001	3.367
1	1	100	0.01500	2.525
	10	1,000	0.01050	2.649
	100	10,000	0.01005	2.664
	∞	∞	0.01	2.665
10	1	10	0.14925	1.557
	10	100	0.10495	1.732
	100	1000	0.10050	1.753
	∞	∞	0.1	1.755

Differences are only great if the population size is very small (10) while selecting only a single line.

6. Variance in selection intensity

If 10 breeders were each given a different random sample of 10 lines from a cross and each selected the best line they could find, there would be considerable variation in the outcomes.

Such variability can be easily observed in **Table 3**: one of the breeders may have been able to identify the best line in 100 (i = 2.525) whereas on average, across all breeders, the expected (i.e. average) selection intensity is only 1.557. There may be other reasons why one breeder performs better than the others, but sampling variation alone has a substantial effect.

There are several methods of calculating the variance of i, but the easiest approach is to use simulation. **Figure 2** shows results from 1,000,000 simulations of selection of the best line from 100.

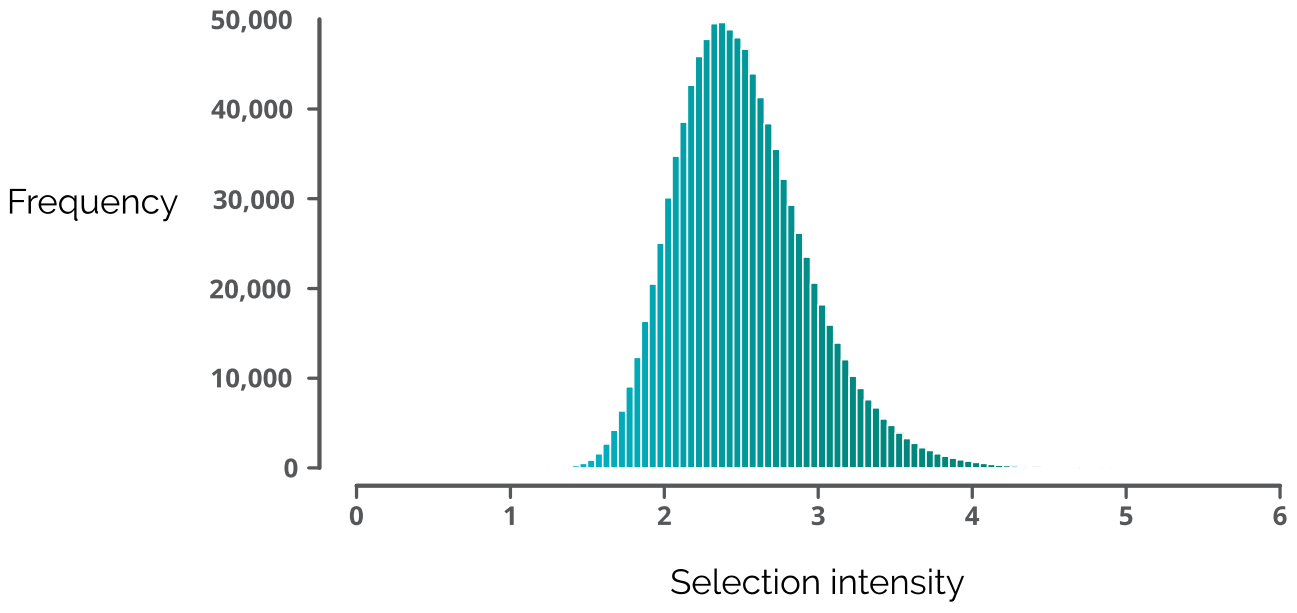

Figure 2. Distribution of i when selecting the best 1 from 100 lines: 1,000,000 simulations.

There is considerable variation in this extreme example. To reduce the variation in selection intensity, the populations size and therefore the number kept after selection must be increased.

Selecting 10 lines from 1000 gives the distribution shown in **Figure 3**, plotted on the same scale.

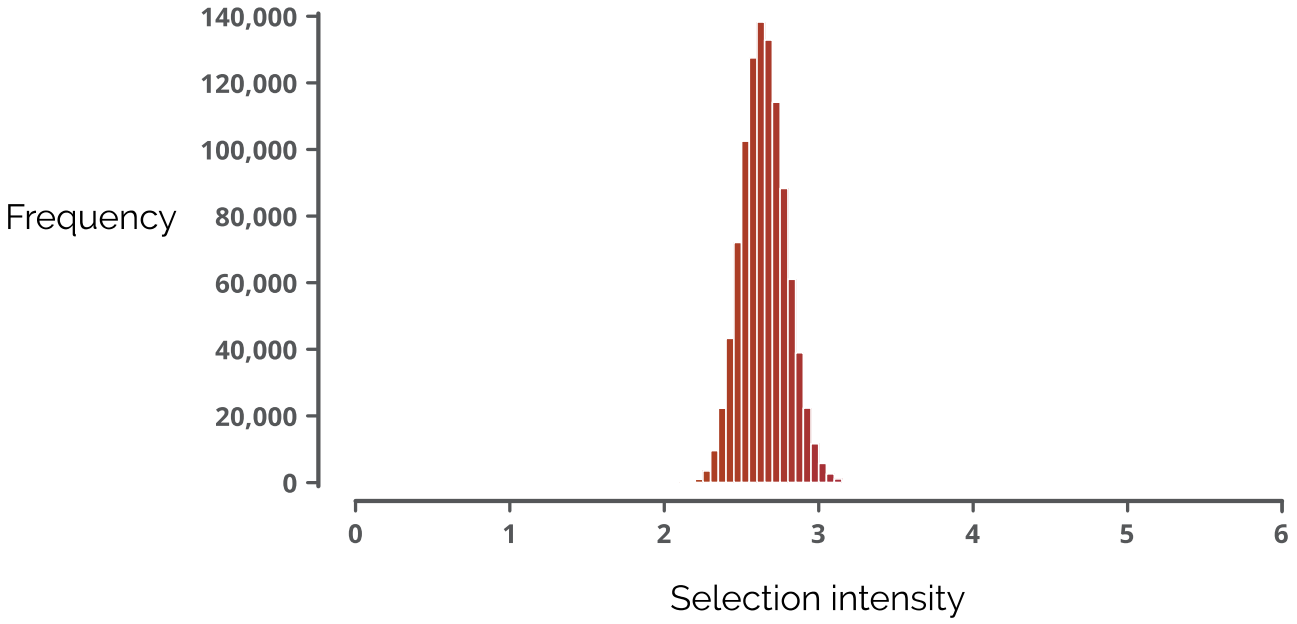

Figure 3. Distribution of i when selecting the best 10 from 1000 lines: 1,000,000 simulations.

The standard deviation of selection intensity is reduced from 0.43 to 0.14. Reliability of response to selection should also be considered when optimizing a breeding program.

7. Alternative selection criteria

It is not always the case that a constant proportion of individuals is selected in each cycle of selection. A common alternative is to select those lines which exceed an individual performance threshold. That threshold could be predefined, for example, a dry matter content of 15%, or it could be determined experimentally: for example, any lines which exceed the performance of the best control.

When population sizes are large enough, selecting against a predefined individual threshold is equivalent to selecting a population proportion threshold: for example, selecting the top 2.5% of a population should be equivalent to selecting any line which exceeds a score of 1.96 standard deviations of the population mean. This is only true in large samples, however. In smaller samples, the proportion of lines that exceed a threshold can vary substantially from sample to sample, while the mean and variance used to define the threshold is usually estimated from the sample itself, and therefore will also vary. In practice, this complication is ignored and is likely inconsequential, but it is worth noting. While it may not be optimal to select a fixed proportion every cycle, this method is simpler and the difference is likely to be slight.

While it is common practice to select lines that exceed the performance of elite control varieties, this compounds the issues of the predefined selection method by adding the factor of error in determining control performance. In practice, the controls should be assessed with greater accuracy, most simply with more replication, with selection among candidates based on best linear unbiased predictions (BLUPs) of candidates compared to best linear unbiased estimations (BLUEs) of controls. If this is done, it is quite possible to find that no lines are selected, but this may be the most realistic outcome.

Finally, selection may not be against a threshold. Rather than all selected individuals contributing equal numbers of progeny to the next generation, some may contribute more. In optimum contribution theory, for example, the best individuals contribute a greater number of progeny to the next generation while a selection of lower ranking individuals will also make a smaller contribution, thereby sacrificing some immediate genetic gain to maintain genetic variation and increase gain in the long term. In this case the selection intensity is the mean of the standardized phenotype, with weights proportional to the contribution of that individual to the next generation.

To compute this without reference to observed trait values, for example when designing a breeding program, ranked normal deviates are required; these are often required for QQ plots, also. Ranked normal deviates can be calculated from the properties of the normal distribution, with an adjustment for small numbers. Most simply, this is done by calculating the quantiles (x-values) of the normal distribution for a nominal proportion selected, running from 0.5/n to 0.95/n where n is the population size.

For example, to calculate the ranked normal deviates from a sample of 10:

n	1	2	3	4	5	6	7	8	9	10
p	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
quantile	1.64	1.04	0.67	0.39	0.13	-0.13	-0.39	-0.67	-1.04	-1.64

In R, qnorm(1-c(0.5:10)/10) will return these values. They are adequate but become less accurate in the extremes. Greater accuracy is provided by simulation, as in the following example featuring 100,000 draws of 10, or from more complex adjustments in the spreadsheet "calc ranked normal deviates":

n	1	2	3	4	5	6	7	8	9	10
simulated	1.54	1	0.66	0.38	0.12	-0.12	-0.38	-0.66	-1	-1.54
accurate	1.54	1	0.66	0.38	0.12	-0.12	-0.38	-0.66	-1	-1.54

If the top three individuals were selected but the best individual contributed twice as many progeny as the second best, the selection intensity would be: 1.54 /2 + 1/4 + 0.66/4 = 1.185

The spreadsheet is annotated and should be easy to use. It relies on a method given in Harter (1961), which also gives the exact solution.

$$\frac{n}{(n-p)!p!}\int_{-\infty}^{\infty}\alpha^{n-p}(1-\alpha)^{p}dx$$

Where:

$\alpha = \int_{-\infty}^{x}\emptyset(x)dx\ \alpha = \int_{-\infty}^{x}\emptyset(x)dx$

$\emptyset(x)$ is the probability density function of x.

Stick to the spreadsheet, or R!

8. Practical considerations

In spite of the diminishing returns from selecting ever harder, high intensities of selection are always better than low. As a thought exercise, consider how low selection can be: as it is never worthwhile to select an individual that ranks lower than the median (random selection without any trait or marker would be better), the lower limit for selection must be 50%.

Theory on selection limits (described in Falconer and Mackay, 1996) also indicates that genetic improvement in the long term - until a selection limit is reached – is maximized by selecting 50% per generation. The basic principle is that we wish to minimize the loss of favorable alleles through drift and maximize their fixation (or maintenance) through selection. This is dependent on population size *after* selection. Optimal contribution theory, which is described in a separate guide, attempts to reconcile the conflict between selection intensity and loss of favorable genetic variation through drift (commonly formulated as increased rates of inbreeding in animal breeding and most of the animal breeding literature).

With finite resources, increasing selection intensity will almost always have a detrimental impact on the other factors in the breeders' equation. In our previous example, for sustained or long-term selection response, selecting 10 individuals generally translates to selecting 1% to 10% of the population (between 100 and 1000 lines before selection), which is acceptable. More intensive selection would require a very cheap phenotyping platform (or genotyping for genomic selection). It is usually worthwhile to increase selection speed at the cost of intensity, even with very high selected proportions (for example 20%, although never higher than 50%).

In single plant selection, it is possible to apply high selection intensity, as the cost of phenotyping (often by visual assessment) is cheap. This is only suitable for highly heritable traits. For yield, heritability is usually low, and high selection intensity does not compensate for this: it is usually more efficient to increase heritability through replication and plot trials, even though selection intensity will be reduced.

Genomic selection implies that large populations can be raised with high selection intensity, and may be justified for this reason alone, as long as prediction accuracy is high enough. If prediction accuracy is low, it may be more efficient to select on phenotype with a lower selection intensity, but greater accuracy (i.e. higher heritability). Genomic selection may provide other benefits, however, such as by reducing cycle time.

In practice, most breeding programs should be built around a long-term population improvement component and a short-term product development component. Optimal selection intensity will be different for each: in the short term, there is no immediate benefit to maintaining genetic variation, the goal is usually to select the best individual line and therefore selection intensity should be reduced. Short-term success is often vital to long-term funding of a breeding program, and it is an error to select with low intensity for short term goals. Such tradeoffs and considerations are best studied through computer simulation.

9. References

1. Bulmer, M.G., 1980. The mathematical theory of quantitative genetics. Oxford, UK: Clarendon Press
2. Falconer, D.S. and Mackay, T.F.C., 1996. Introduction to quantitative genetics. Essex. UK: Longman Group
3. Harter, H.L., 1961. Expected values of normal order statistics. Biometrika, 48(1-2), pp.151-165.