***Exploring the impact of sampling***

***Goal:*** Real studies encounter a variety of issues that affects the conclusions that can be drawn from data analysis. Here, we explore the real world problem of biases in sampling and what happens when the researcher does not know about the reason for the bias. We also then investigate how the results from data analysis improve as the researcher knows more about the underlying mechanisms leading to biased sampling.

***Step 1. Sampling and non-stochastic environment***

**Sub-goal**: to explore how hidden patterns in environment combined with variance in sampling affect estimates of variance parameters and their interpretation.

**Introduction**: In the previous module (*Basic Lessons about Variance*), we partitioned phenotypic variance into several components (the variance among individuals, VI, the variance caused by measurement error, Vm, and variance caused by the environment, VE). In the final step of that module, we illustrated how measurement of the environment could help explain some of the variance. Often, when we study phenotypes in natural populations, many aspects of the environment that could affect phenotypes will be unknown and so not measured. In Step 3, this unmeasured environmental variance ended up as “residual” variance, and it had no effect on the estimate of among-individual variance because the environment was randomly determined from one sampling period to another and all individuals were sampled at the same time and experienced the same environment. In the present module, we explore what happens when we relax this obviously simplified assumption. For example, suppose the environment changes steadily over the sampling period. What happens when the pattern of how an investigator measures individuals varies, such as if the timing of measurement is different for different individuals?

**Exercise**: As in previous simulations, we will generate a new group of individuals, with phenotypic variation generated by measurement error (Vm), individual differences (VI), and the impact of a specified environmental variable x which produces variance due to the environment (VE). We have shifted to using the notation VE here instead of which we used in Step 3 and 4 of the module “*Basic Lessons about Variance*”. We do this because we will soon explore what happens when only some of the environmental variance is known, and we will use for that known variance.

As before, you can set Vm, VI and VE, and for this module, VE must be greater than 0.

Note that from now on, the total variance (Vp) is not restrained to 1 anymore and the proportion of each variance component is shown next to the input element.

Also, the number of individuals will be set to 100 all along this module.

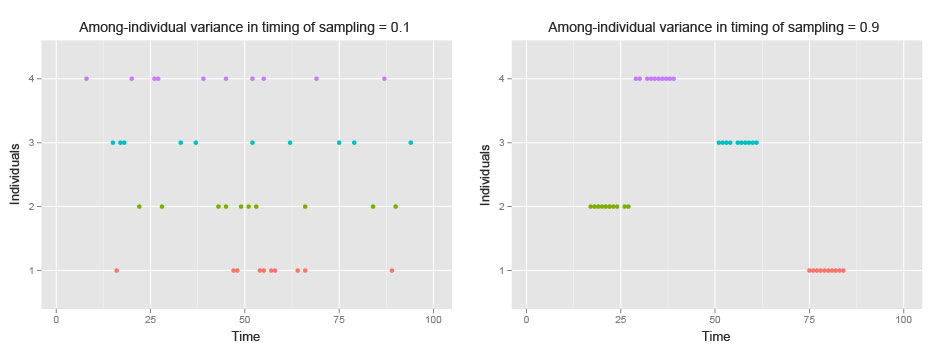
<Input Vm, VI >

<Input environmental effect variance, VE>

The environment for this simulation is, for convenience, set as being linear over time, affecting all individuals similarly (i.e., it is “shared”). The environment is also expressed in unit variance (i.e., Var(x)=1) and mean-centered (i.e., E(x)=0).

You also must enter parameters for variance in the sampling timing within and among individuals. For this simulation, the total number of expressions of the phenotype from which you can sample is fixed at 100. While you can vary the number of individual samples taken, for this module to effectively illustrate the issues with sampling, the number of samples must be much less than 100. The key parameter to be entered by you will be the among-individual variance in timing of those records. To illustrate, below are examples of sampling records for a small number of individuals when the among-individual variance in sampling timing is 0, and when it is 0.9.

< Provide graphs of ~5 individuals (color-coded) through time to illustrate effect of sampling variance in timing.>



Now you can input your own values.

<Input number of records taken, and among-individual variance in timing of those records>

The figure below shows time of sampling of a subset of individuals according to the values entered.

<Immediately show graphical output of subset of individuals illustrating sampling for the values user entered.>

< Run simulation>

**Results**

If we have no information about the environment, the model we incorrectly assume to be true is:

**R code:**

# install.packages("lme4")

LMM <- lme4::lmer(Phenotype ~ 0 + (1|Individual), data = sampled\_data)

A mixed-effects statistical model can then estimate these model parameters:

Statistical output:

|  |  |
| --- | --- |
| True | Estimated |
| Individual Variance (VI) = …… | Individual variance in sample (V’I) = ….. |
| Measurement error variance (VM) = ….. | Residual variance of sample (V’R) = ….. |
| Environmental effect variance (VE) = ….. |

The above should show that if the unmeasured environment changes over time AND there is among-individual variance in sampling, then some of the unknown VE is placed into residual variance (making residual variance larger than just measurement variance Vm), and some ends up in the estimated VI, also making it bigger than it should be.

**Conclusion**

This exercise demonstrates that if there is among-individual variance in timing of sampling, then estimates of VI will be incorrect since inevitably there are systematic differences in environments over time. Sampling biases thus can produce “pseudo-personality” or “pseudo-repeatability” (see also Dingemanse and Dochtermann 2013) and could mislead a researcher into believing there are consistent differences between individuals when there are none (or they are much smaller than it appears).

Because among-individual variance in sampling and systematic changes in environment are extremely likely in real systems, how can we get accurate estimates of VI?

We explore two solutions to this problem:

1. Adjust sampling regime to minimize it (go to Step 2)
2. Accounting for biases in your analysis (go to Step 3).

***Step 2. Sampling to reduce effects of non-stochastic environment***

**Sub-goal**: Using simulations to generate sampling regimes that limit the effects of non-stochastic environments.

**Introduction**: Step 1 revealed a problem—non-stochastic environments through time and variability in the timing of sampling can create biases in estimates of among-individual variation. In this step we encourage you to adjust the sampling regime to minimize this problem. It should be obvious that if all individuals are sampled with the same timing, then the bias in the estimates of among-individual variance disappears, but it is worthwhile assessing how close one has to be to identical sampling and whether there are biases in other parameters that remain. So, in this step we will allow you to simulate several types of non-stochastic environments and adjust the sampling regime.

**Exercise**: As in Step 1, we will generate a new group of individuals, with phenotypic variance caused by measurement error (Vm), individual differences (VI), and the impact of the environment (VE).

<Input Vm, VI and VE>

You now get to set the environment. In Step 1 of this module, we used an environment that was experienced similarly by all individuals (“shared”) and which changed systematically over time. Below, you can change these settings to have environments that each individual experiences uniquely (“unshared”), and which changes over time as some other function (e.g., stochastically or as a regressive autocorrelated decay function).

<Input environmental effects: shared or unshared; stochastic, linear, or auto-regressive decay>

As in step 1, you also must enter parameters for variance in the sampling timing within and among individuals. As before, the number of expressions of the phenotype will be set by us at 100, so keep this in mind as you enter values here.

<Input number of records taken, and among-individual variance in timing of those records>

Immediately show graphical output of subset of individuals illustrating sampling for the values user entered.

< Run simulation>

**Results**

As before, the model we assume to be true (but which is not since the environmental effect is not included) is:

**R code:**

# install.packages("lme4")

LMM <- lme4::lmer(Phenotype ~ 0 + (1|Individual), data = sampled\_data)

A mixed statistical model estimates the parameters which we can compare with the true values:

Statistical output:

|  |  |
| --- | --- |
| True | Estimated |
| Individual Variance (VI) = …… | Individual variance in sample (V’I) = ….. |
| Measurement error variance (Vm) = ….. | Residual variance of sample (V’R) = ….. |
| Environmental effect variance (VE) = ….. |  |

**Conclusion**

The results of any given simulation may vary, but the overall picture that emerges if you do several simulations should be that your estimates are better when VE is small, as you measure each individual more often, and your sampling time is increasingly similar among individuals.

Did you simulate a population where the environment is not shared among individuals? If not, try it now. What you should find is that no matter what the sampling regime, your estimate of VI is too high. To understand, let’s return to the definitions of the variance components: We defined VI as the variance among individuals that permanently affected their phenotype throughout the sampling period. Biologically, this can be ascribed to genetic differences or environments acting during development (e.g., before measurements started). When environments are unshared during sampling, the environment is affecting the phenotype each time it is expressed. However, because the environment is autocorrelated across sampling episodes and differs among individuals, apparent individual differences arise because individuals are in different environments not because they entered the time period of phenotypic expression differing in their phenotype (note: You may be thinking that since individuals in the real world partially choose their environment then their phenotype is not solely due to the environment. That is true but does not change the fact that for the focal trait it is sensitive to the environment the individual is in each time it is expressed. We will get to the issue of multiple phenotypic characters and how they might integrate in the “*multiple traits*” module).

To conclude for this step, if you do not know what environments are affecting trait expression, sampling in parallel for all individuals is a possible solution to potential biases created by non-stochastic environments. But, because unshared environments can create biases even with identical sampling (and often identical sampling will be nearly impossible to achieve), the only other solution is to measure the environment and account for possible biases explicitly. This is explored next in Step 3.

***Step 3. Biased sampling and known and unknown environments***

**Sub-goal**: Accounting for the environment to control for environmental biases.

**Introduction**: Step 1 of this module illustrated that environmental effects on phenotypes can produce biases in estimates of among-individual variance (VI). Step 2 explored how altering sampling regimes could reduce this problem but also revealed that in some circumstances no sampling regime would work. Sometimes individuals experience different environments, and no sampling regime can adjust for that. However, if investigators can measure the environment, then such differences could be accounted for. Environmental variance was accounted for using linear regression in step 4 of the “*Basic Lessons about Variance*” module. Here we demonstrate that this can, under some circumstances, solve the bias in sampling problem.

**Exercise 1**: This exercise follows the same structure as all of our other simulations so far. We will generate a group of individuals, with phenotypic variance caused by measurement error (Vm), individual differences (VI), and the impact of the environment (VE). So, first set the true values of these variances:

<Input Vm , VI and VE>

The environment can be chosen as in Step 2. It, combined with the sampling regime, will affect within- and among-individual variance in the environment.

<Input environmental effects: shared or unshared; stochastic or autocorrelated or linear or cyclic>

Finally, we will have you set how much of the environmental variance has been measured and is therefore known. You will select a proportion, from 0 to 1 of this variance. This proportion along with the proportion of total variance that is environmental will determine the correlation between phenotype and the known environment. The results of Step 1 should have shown you what happens when all the environmental variance is unknown (or not included in your statistical model). Here, let’s start with all the environmental variance being known and measurable.

< Run simulation>

**Results**: In the module “*Basic Lessons about Variance*”, Step 4, we said the statistical model was

**R code:**

# install.packages("lme4")

LMM <- lme4::lmer(Phenotype ~ 1 + X1 + (1|Individual), data = sampled\_data)

This is the model we will investigate here. We will compare it to a model in which all of the environmental variance is unknown, e.g.,

**R code:**

LMM <- lme4::lmer(Phenotype ~ 1 + (1|Individual), data = sampled\_data)

A mixed effects statistical model estimates the parameters, which we can compare with the true values:

Statistical output:

|  |  |  |
| --- | --- | --- |
| True | Totally unknown environment | Environment known (proportion = 1.0) |
| Population intercept (β0) = 0 | Population estimated mean (β’0) = … | Population estimated mean (β’0) = … |
| Individual Variance (VI) = …… | Individual variance (V’I) = ….. | Individual variance (V’I) = ….. |
| Measurement error variance (Vm) = ….. | Residual variance (V’R) = …. | Residual variance (V’R) = …. |
| Environmental effect variance (VE) = ….. |  | Estimate of known environmental variance () = ….. |
| β = ….. |  | β’ = ….. |

This should show you that when there is among-individual variance in sampling and you can account for all the environmental variance with an x variable, any bias in VI caused by the biased sampling disappears.

A brief reminder about notation: When unknown environments affect phenotypic variance, we have referred to that variance as VE. In the model where the environment is known (x), there now is a specific component of variance due to that known environmental factor, . In the case above, VE = , but in the real world with many environmental variables, will be only a fraction of VE.

**Exercise 2:** Now, let’s repeat the same simulation as above, expect this time explore what happens as you change the proportion of the environmental variance that is known. Below is the bar that allows you to adjust this.

<Enter proportion of environment that is known (between 0 and 1)>

If you want, you can also change the level of bias in sampling.

<Reshow bar for bias in sampling with counter set on previous value; allow user to change it or not>

RUN

**Results**: As above, we will show you the true values you entered, the values estimated when the environment is unknown, and those estimated when some portion of the environment is known and included in the model.

Statistical output:

|  |  |  |
| --- | --- | --- |
| True | Totally unknown environment | Environment known (proportion = …) |
| Population intercept (β0) = 0 | Population estimated mean (β’0) = … | Population estimated mean (β’0) = … |
| Individual Variance (VI) = …… | Individual variance (V’I) = ….. | Individual variance (V’I) = ….. |
| Measurement error variance (Vm) = ….. | Residual variance (V’R) = …. | Residual variance (V’R) = …. |
| Environmental effect variance (VE) = ….. |  | Estimate of known environmental variance () = ….. |
| β = ….. |  | β’ = ….. |

**Conclusion:** There are two lessons that emerge from this exercise. First, biases in sampling are usually inevitable, but measuring the underlying environments that differ among individuals can reduce them. Thus, if you want to measure among-individual variance, you must think carefully about potential biases in environments, and measure those environments. That will give you a better estimate of among-individual variance.

The second lesson is that bias in sampling may occur without you being aware of it. This unknown environment will affect your estimate of among-individual variance. Put another way, any among-individual variance estimated from real data could be due to unknown biased environments. One cannot be sure that you have accounted for all of the environmental variance. The VI that is found from real data must therefore be interpreted cautiously.

**A final caveat:** An interesting consequence of having variance in sampling among individuals is that it produces variance in the experienced environment that exists both within and among individuals. We have assumed that the impact of the environmental variance that exists among individuals is the same as that of the variance in environment within-individuals. As an example, individuals may be on territories with different average levels of resources through the whole period of time you are taking measurements, and those resources may fluctuate some from day to day as well. Thus in your population, there is both among-individual variance in environment (e.g., differences between territories) and within-individual variance in environment (differences between days within a territory). We have assumed these have the same effect on phenotype. It is possible that this is not the case. If so, the method we have demonstrated here will not give accurate estimates of VI. We discuss one solution to this in a module on within and among-subject centering. The issues related to centering are complex, so we recommend this module be done after the module on random regression.