Bits of Information to help you understand evolutionary rates

Ian Dworkin

2025-11-27

This is a disorganized tutorial on some numeric and statistical issues to help you understand some technical issues at play in our discussion of evolutionary rates. Importantly all of these are useful much more broadly and come up in data modeling all the time.

Setting things up

Usually with evolutionary rates we start with two things a measure of phenotypic or molecular change across two or more time points, and some measure of the time intervals at which these changes occur.

So we can call our variable containing information about the change in phenotypes d for "difference" or "divergence". For time we can start with some basic time interval over which we are measuring the differences in phenotypes. For the moment, I will just call this variable τ (pronounced tau) for this time interval we are measuring on. Why not just t? Well in R `t` has an explicit meaning (transpose of a matrix), which we are not interested in. Also more generally, the time intervals we are considering can vary in all sorts of ways.

At the most basic we can define an evolutionary rate ρ ("rho") as the amount of change per some unit of time:

$$\rho = \frac{d}{\tau}$$

Scale matters, but it can be complicated

Hopefully this all seems simple enough. We want to *scale* the amount of change in phenotype, d, to enable comparisons. For instance what if we wanted to compare the amount of change in wing length in *Drosophila melanogaster* after 100 generations of experimental evolution at high temperature to a similar experiment conducted in a mosquito, *Culex pipiens*, that was done for 20 generations of experimental evolution.

In the case of *D. melanogaster*, the average wing length at the beginning of the experiment was $\bar{z}_{Dm,0} = 1.0\ mm$, and after evolving for 100 generations at high temperatures, the average wing length in the population was $\bar{z}_{Dm,100} = 0.7\ mm$. In *C. pipiens* the average length was $\bar{z}_{Cp,0} = 4.0\ mm$ at the beginning of the experiment and $\bar{z}_{Cp,20} = 3.75\ mm$

We could just compare the **amount** of change observed in each species $d_{Dm,100}$ to $d_{Cp,20}$.

$$d_{Dm,100} = |\bar{z}_{Dm,0} - \bar{z}_{Dm,100}| = |1\ mm - 0.7\ mm| = 0.30\ mm$$
$$d_{Cp,20} = |\bar{z}_{Cp,0} - \bar{z}_{Cp,20}| = |4.0\ mm - 3.75\ mm| = 0.25\ mm$$

In R, we would write this as so:

```r
d_dm = abs(1.0 - 0.7)

d_cp = abs(4.0 - 3.75)
```

```r
d_dm
```

```
## [1] 0.3
```

```r
d_cp
```

```
## [1] 0.25
```

So the absolute amount of change between the *Drosophila populations* seems greater that in the mosquito, right?

Can you think of any problems with this comparison? Don't scroll to the next page until you have thought about this first!!!

Rates make the amount of change easier to compare

Even *if* there were similar selection pressures and genetic variance for wing length in both species, in one experiment there were 20 generations of opportunity for evolutionary change, and in the other experiment there were 100 generations. So we may wish to use a rate to assess the amount of change **per generation**.

$$\rho_{Dm} = \frac{d_{Dm,100}}{\tau_{Dm}} = \frac{0.3\ mm}{100 - 0} = 0.003\ \frac{mm}{generation}$$

```
rho_dm <- 0.3/100
rho_dm
```

```
## [1] 0.003
```

and in the mosquito

$$\rho_{Cp} = \frac{d_{Cp,20}}{\tau_{Cp}} = \frac{0.25\ mm}{20 - 0} = 0.0125\ \frac{mm}{generation}$$

```
rho_cp <- 0.25/20
rho_cp
```

```
## [1] 0.0125
```

So now it seems like the rate of change of wing length is greater in the mosquito than in *Drosophila*, even though the *total* amount of evolutionary change was greater in *Drosophila* (after 100 generations) as compared to mosquitos (after just 20 generations).

Can you think of any potential issues with this comparison? Spend some time before going to the next page.

Scale matters... Think proportionally!

One major issue we need to consider is the fact that the mosquitoes are much bigger than *Drosophila*. So the larger amount of change in wing length may just be a simple consequence of the scale. Thing about this for a more extreme example. A 1mm difference in human height is basically rounding error, while a 1mm difference in the *Drosophila* wing length means a wing that is twice as long!

So we want to think about proportional differences. So we are going to scale the amount of change we observe by some "ruler" (that will serve as our denominator for the proportion). There are many common choices for such rulers, and a few have been used in evolutionary rates. To start with though, we will just scale the amount of phenotypic change in wing length by the mean wing length in the ancestral population. We would do this for each species.

$$\rho_{Dm} = \frac{d_{Dm,100}}{\tau_{Dm}} = \frac{\frac{0.3\ mm}{1\ mm}}{100 - 0} = 0.003\ generation^{-1}$$

```
rho_dm <- (0.3/1.0)/100
rho_dm
```

```
## [1] 0.003
```

and

$$\rho_{Cp} = \frac{d_{Cp,20}}{\tau_{Cp}} = \frac{\frac{0.25\ mm}{4\ mm}}{20 - 0} = 0.0031\ generation^{-1}$$

```
rho_cp <- (0.25/4)/20
rho_cp
```

```
## [1] 0.00313
```

Now we have measures of evolutionary change where we are examining proportional changes relative to the mean wing size of the ancestral populations. We see that the proportional amount of change in wing length per generation is actually pretty similar across the two species.

Log transformations help to make you focus on proportional changes

The easiest way of making your focus be on proportional changes is to log transform your data. For a lot of the data types we work with in biology log transformations are a sensible default anyways (and about 2% of the time they create some issues, but ignore that for the moment). Let's just log transform (natural log is a good default) our original values for wing sizes.

So instead of (for Drosophila)

```
d_dm <- 1.0 - 0.70
d_dm # difference on arithmetic scale
```

```
## [1] 0.3
```

We would use:

```
log_d_dm <- log(1.0) - log(0.70)
log_d_dm # difference on log scale
```

```
## [1] 0.357
```

and for the mosquito:

```
log_d_cp <- log(4.0) - log(3.75)
d_cp # Difference on arithmetic scale
```

```
## [1] 0.25
```

```
log_d_cp # difference on log scale
```

```
## [1] 0.0645
```

Immediately it becomes apparent that the proportional amount of phenotypic change (across the entire time period, not scaled to number of generations) is greater in Drosophila than in mosquitoes.

This will carry through when we convert these to rates (per generation)

For Drosophila:

```
rho_d_dm_V2 <- log_d_dm/100
```

For Mosquitoes:

```
rho_d_cp_V2 <- log_d_cp/20
```

```
rho_d_dm_V2
```

```
## [1] 0.00357
```

```
rho_d_cp_V2
```

```
## [1] 0.00323
```

You can see we get to a pretty similar place in terms of our estimated rates simply by using log transformed values of our data. In practice, for small amounts of change the natural log transformed data approximates what we expect if we scale by the mean. This is particularly useful when we are interested in variation for traits.

CV and standard deviation of log transformed variables

While I won't go into it in detail here, from the above comparison of mean scaled and log transformed values you can get a sense that even when we are looking at variation in a trait that log transformation can be very helpful. Indeed it turns out that for many biologically relevant variables the standard deviation estimated on natural log transformed values approximates the *coefficient of variation* which is the mean scaled standard deviation. i.e.

$$CV_x = \frac{\hat{\sigma_x}}{\hat{\mu}_x}$$

With $\hat{\sigma_x}$ being the estimated *standard deviation* for your variable (like wing length) x. $\hat{\mu}_x$ is the estimated *mean* of x.

If we log transform x, then it turns out

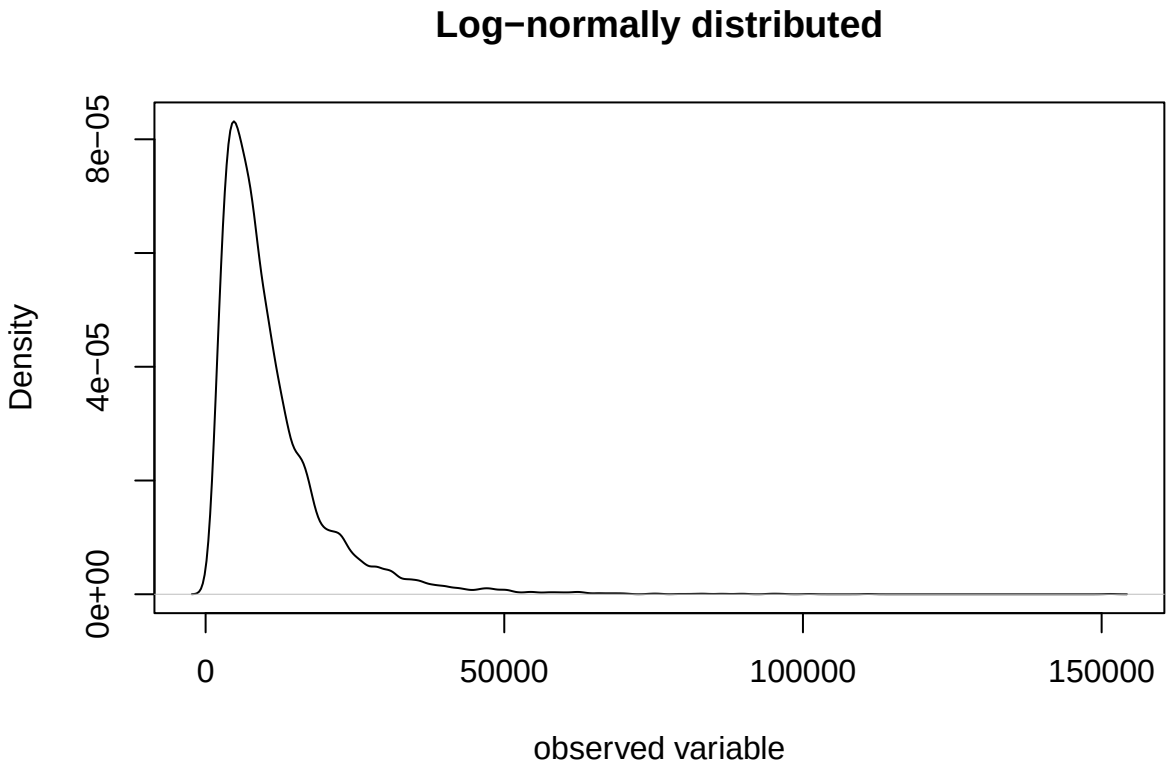
$$\hat{\sigma}_{log(x)} \approx CV_x$$

So once again, log transformation can be very helpful.

log normal is the norm for many biological traits

While this is not the right place for going to this in detail, it is worth knowing that many biological variables are *log-normal* distributed, not *normally* (Gaussian) distributed. That is, if you plot the distribution of the observed values you measured, the distribution will tend to look like this:

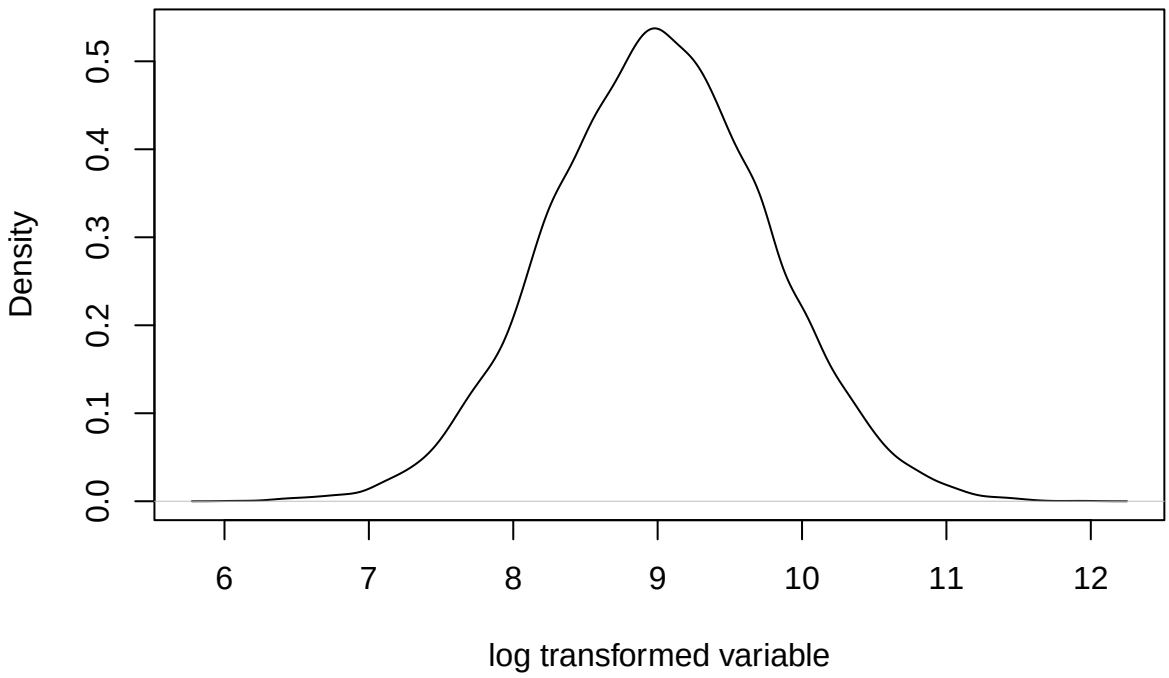
```
dummy_x <- rlnorm(10000, meanlog = 9, sdlog = 0.75)

plot(density(dummy_x), xlab = "observed variable", main = "Log-normally distributed")
```



Which shows the characteristic long "right tail" common for log-normally distributed variables.

If we plot the log transform the variable, we will see it looks much more Gaussian (normally) distributed

```
plot(density(log(dummy_x)), xlab = "log transformed variable", main = "")
```



Take care when comparing traits that have different dimensions (lengths, areas, volumes)

Often in studies of evolutionary rates comparisons are being made across many different types of traits. While the various approaches (and in particular log transformation) can help when the scale is very different (i.e. Drosophila to mosquito to human), we still need to think about the *dimensionality* of the traits we are measuring. For instance for morphology are we measuring a length of an organ like the wing? What happens if we measure the wing area instead?

Let's say we also measured wing width in *Drosophila* at $G_0 = 500\mu m$ and $G_100 = 420\mu m$. We decide to be a bit lazy and assume that wing area can be approximated as an ellipse (bad approximation, but let's use it).

$$A = \pi LW$$

So in R we could calculate this at both G_0 and G_100.

```
dm_A_0 = pi * 1000 * 500
dm_A_100 = pi* 700 * 420
```

```
dm_A_0
```

```
## [1] 1570796
```

```
dm_A_100
```

```
## [1] 923628
```

So the wing Area is $A_{G0} \approxeq 1570796\ \mu m^2$ and $A_{G100} \approxeq 923628\ \mu m^2$. If we log transform these area we get ~ 14.27 and ~ 13.74. Just to remind you we could also do this all in log transformed values from the beginning:

```
log_dm_A_0 = log(pi) + log(1000) + log(500)
log_dm_A_100 = log(pi) +  log(700) + log(420)
```

```
log_dm_A_0
```

```
## [1] 14.3
```

```
log_dm_A_100
```

```
## [1] 13.7
```

So for wing length our evolutionary rate (from log transformed values in μm) would be

```
log_WL_dm_rate <- (log(1000) - log(700))/100
log_WL_dm_rate
```

```
## [1] 0.00357
```

And for wing area

```
log_WA_dm_rate <- (log_dm_A_0 - log_dm_A_100)/100
log_WA_dm_rate
```

```
## [1] 0.00531
```

So does this suggest wing area is evolving faster than wing length?

Can you think of any potential issues with this comparison? Spend some time before going to the next page.

You need to account for the dimensionality of the traits you are measuring

The main issue with the comparison is that area is in (base) units of μm^2 while length is in μm. To make these comparable to one another we would need to take the square root of the area measures $\sqrt{WA_{G0}}$ and $\sqrt{WA_{G100}}$ before calculating the amount of evolutionary change in wing area (if we are comparing to linear measures like length). If we were working with volume we would instead use the cubic root.

Alternatively if we are working on log transformed data we just need to divide our measures by two for area or divide by 3 for volumes.

So our corrected estimate would be

```
log_WA_dm_rate_corrected <- (log_dm_A_0 - log_dm_A_100)/(2*100)
log_WA_dm_rate_corrected
```

```
## [1] 0.00266
```

Which is actually a bit lower evolutionary rate than just for wing length.

Negative correlations are the null expectation when comparing a rate with time!

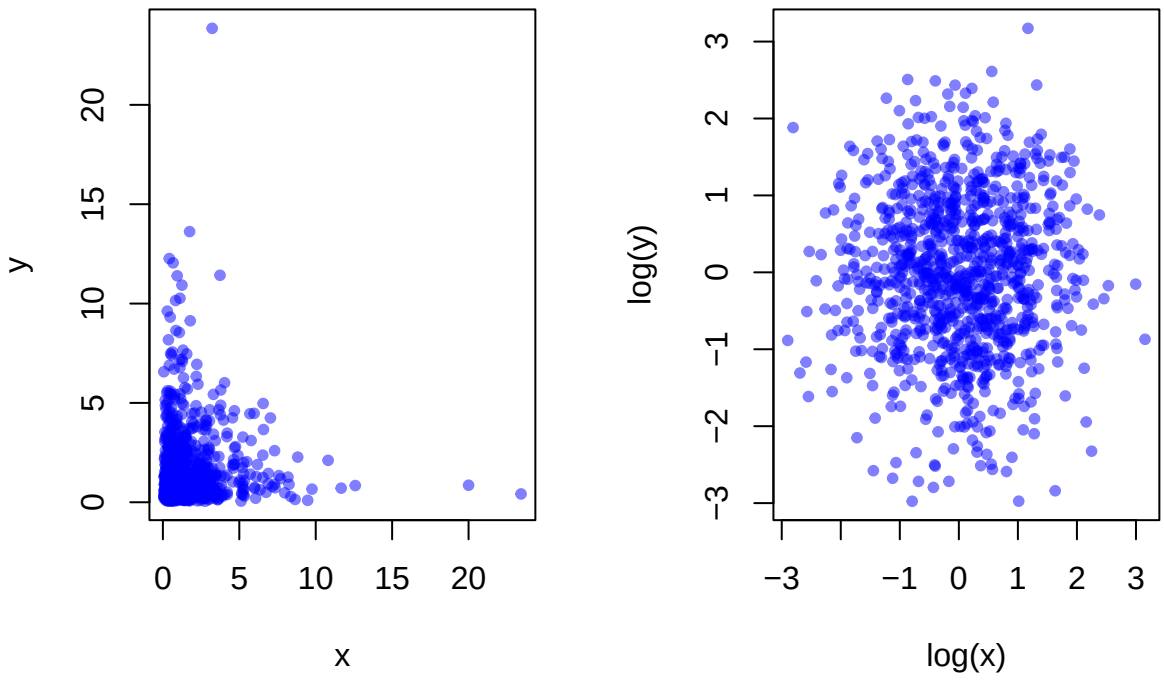
```
x <- rlnorm(1000)

y <- rlnorm(1000)
```

```
cor(x, y)
```

```
## [1] -0.00658
```

```
cor(log(x), log(y))
```

```
## [1] 0.00541
```



but let's say we looked at the ratio of y/x (like we do with rates) and compare it with x.

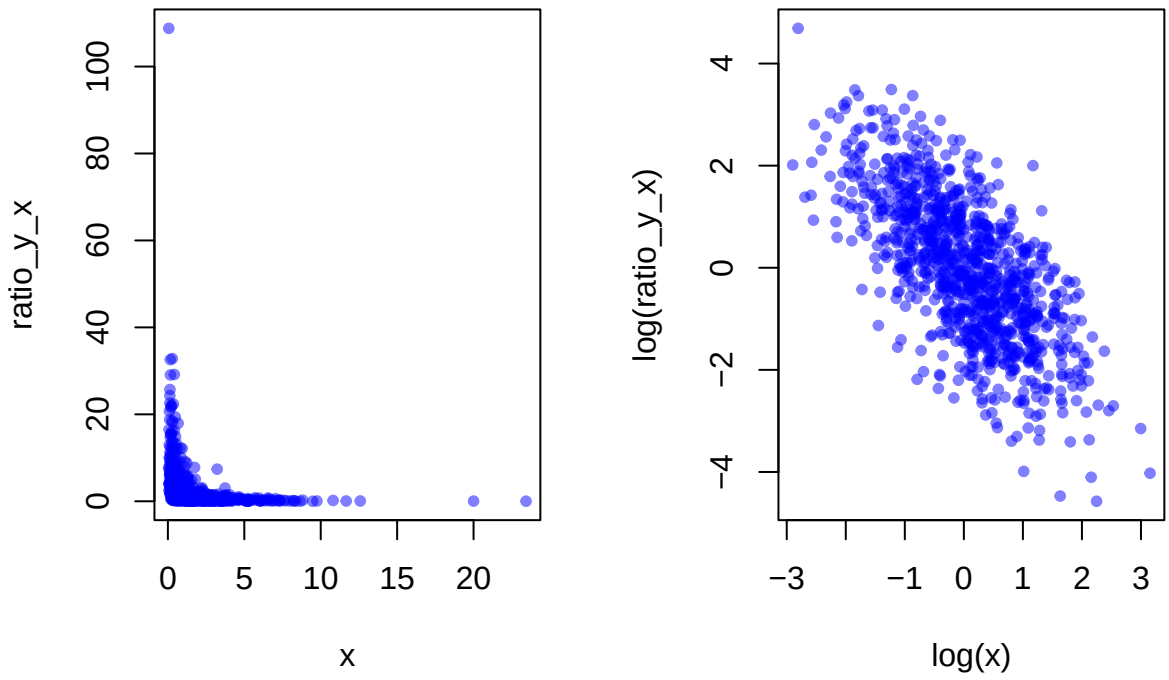
```
ratio_y_x <- y/x
```

```
cor(ratio_y_x, x)
```

```
## [1] -0.248
```

```
cor(log(ratio_y_x), log(x))
```

```
## [1] -0.689
```



With a slope near -1!

```
cov(log(ratio_y_x), log(x))/var(log(x))
```

```
## [1] -0.994
```

This is a critical point, and really relevant to evolutionary rate studies. The null expectation when there is no biological relationship between the amount of phenotypic change and time, is to observe a negative association between rate and time.

Below I have written a little simulator you can use to play with this idea

```
simmie1 <- function(mean_x = 0, sd_x = 1, mean_y = 0, sd_y = 1) {
  x <- rlnorm(n = 1000, meanlog = mean_x, sdlog = sd_x)
  y <- rlnorm(n = 1000, meanlog = mean_y, sdlog = sd_y )

  log_x <- log(x)
  log_y <- log(y)

  y_x_ratio <- log(y/x) # ratio variable we are making

  cov_val <- cov(log_x, log_y) # covariance of x and y (log transformed)
  cor_val <- cor(log_x, log_y) # correlation x and y (log transformed)
  slope_y_x <- cov_val/var(log_x) # slope of y~x

  cov_val_rt <- cov(log_x, y_x_ratio) # covariance of x and ratio
```

```
  cor_val_rt <- cor(log_x, y_x_ratio)
  slope_rt_x <- cov_val_rt/var(log_x)

  return(c(cov_xy = cov_val, cor_xy = cor_val, slope_y_x = slope_y_x,
           cov_rt = cov_val_rt, cor_rt = cor_val_rt, slope_rt = slope_rt_x))
}
```

Test that it works. It should spit out six values

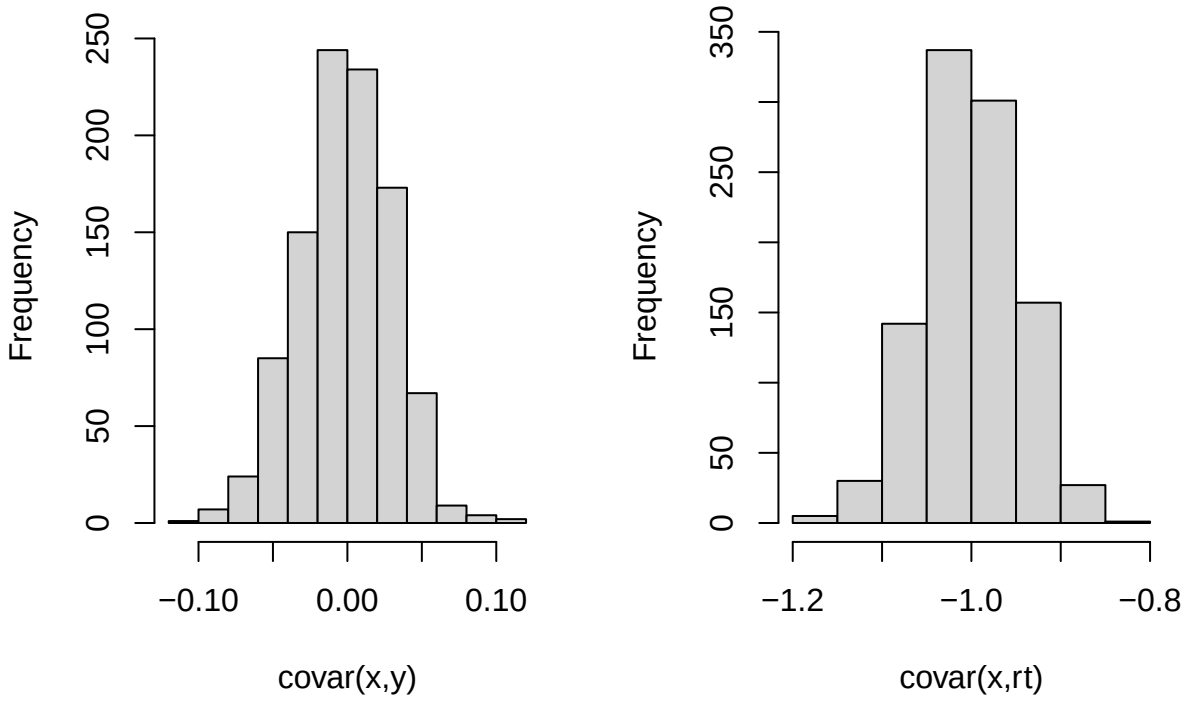
```
simmie1()
```

```
##    cov_xy    cor_xy slope_y_x    cov_rt    cor_rt  slope_rt
##   -0.0517   -0.0514   -0.0501   -1.0836   -0.7332   -1.0501
```

Let's do 1000 iterations of this simulation:

```
simulation_output <- t(replicate(1000, simmie1()))
```

Let's now compare what the distributions of covariances look like for just y vs x, and the ratio variable with x



Important take home point: Even when there is no relationship between the time interval and amount of change in the trait, just comparing a ratio variable $\frac{y}{x} \sim x$ there is a strong negative association.Think about what this might mean when examining change in evolutionary rates with amounts of time!