Wildflower plantings attract more rare species, but I'm not impressed

Michael Roswell

2023-02-22

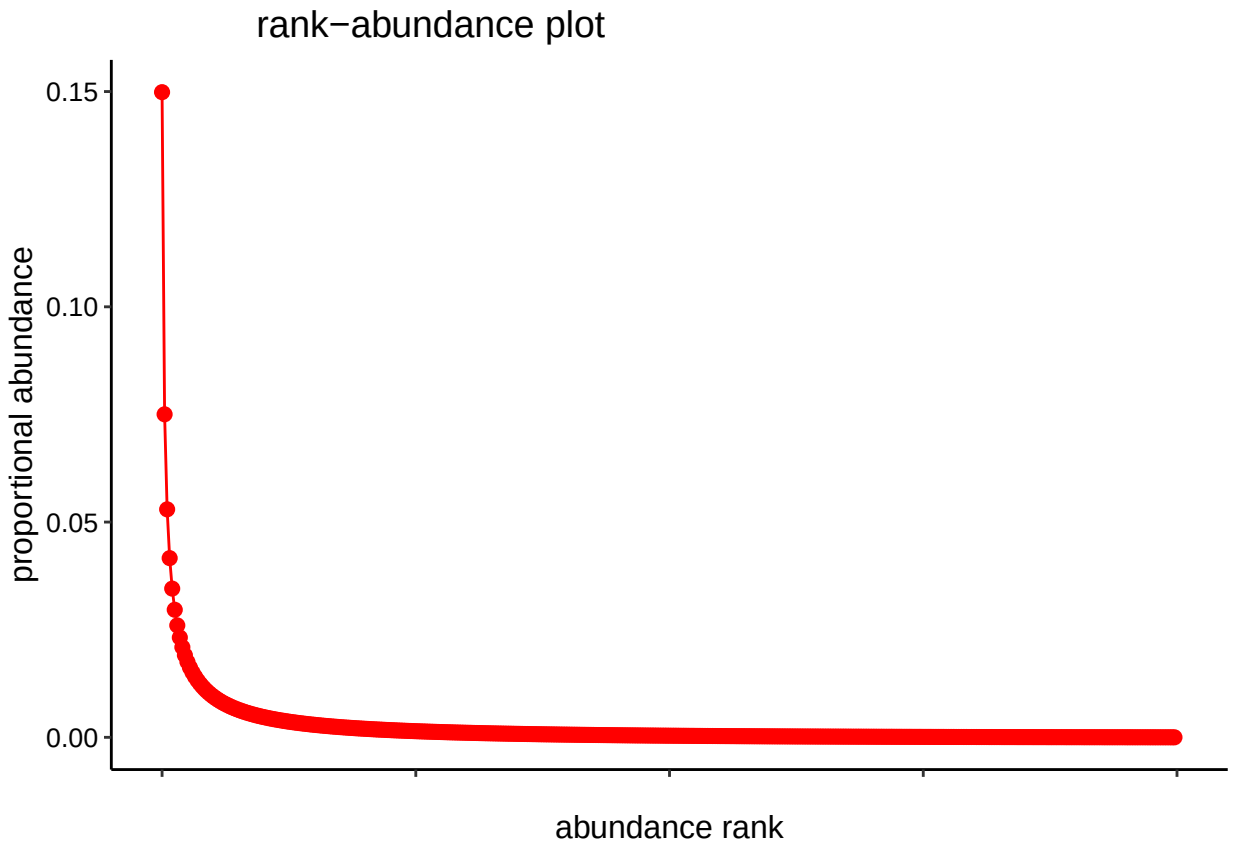
Hook

I'd like to use this as the title for a paper, but I'm struggling to say why I'm not impressed when it looks like rare species increase *more* than common species do in response to a treatment. I think it's not because the treatment is interpreted differently by rare and common taxa; instead I think it's because of something numerical I can't quite pin down... but it's an uninteresting (or at least, neutral) sampling effect, not a niche-based difference in rare and common spp (which I would think is cool!)

Let me show you a sampling effect that I don't think is interesting.

First, let me introduce you to the "bees" in my imaginary study region. There are 400 species, and like any natural species abundance distrbution, there are a few common species and a long tail of very rare ones. I simulated the distribution of relative abundances using the **R** package **MeanRarity**, but honestly don't get hung up on this part, the point is a few species are common and most are rare.

```
rich <- 400
regional_abundances <-
  MeanRarity::fit_SAD(rich = rich, simpson = 25)$rel_abundances
MeanRarity::radplot(regional_abundances) +
  ylab("proportional abundance")
```



Now, I'll define some species as "rare" and others as common. Let's define the top 100 species as common, and the bottom 100 as rare. This is one way to think about commonness and rarity... admittedly not the most typical way but it's clear, right? The top quarter are common, the bottom quarter are rare, if you're in the middle you're neither common nore rare..

Let's say I could sample the species from this regional pool perfectly in proportion to their relative abundance. If in sample "a" I sampled 100 individuals and in sample "b" I sampled 200 individuals, what would happen? I mean, would I see a bigger increase in common species, or in rare species, between "b" and "a"? Why?

Let's check it out.

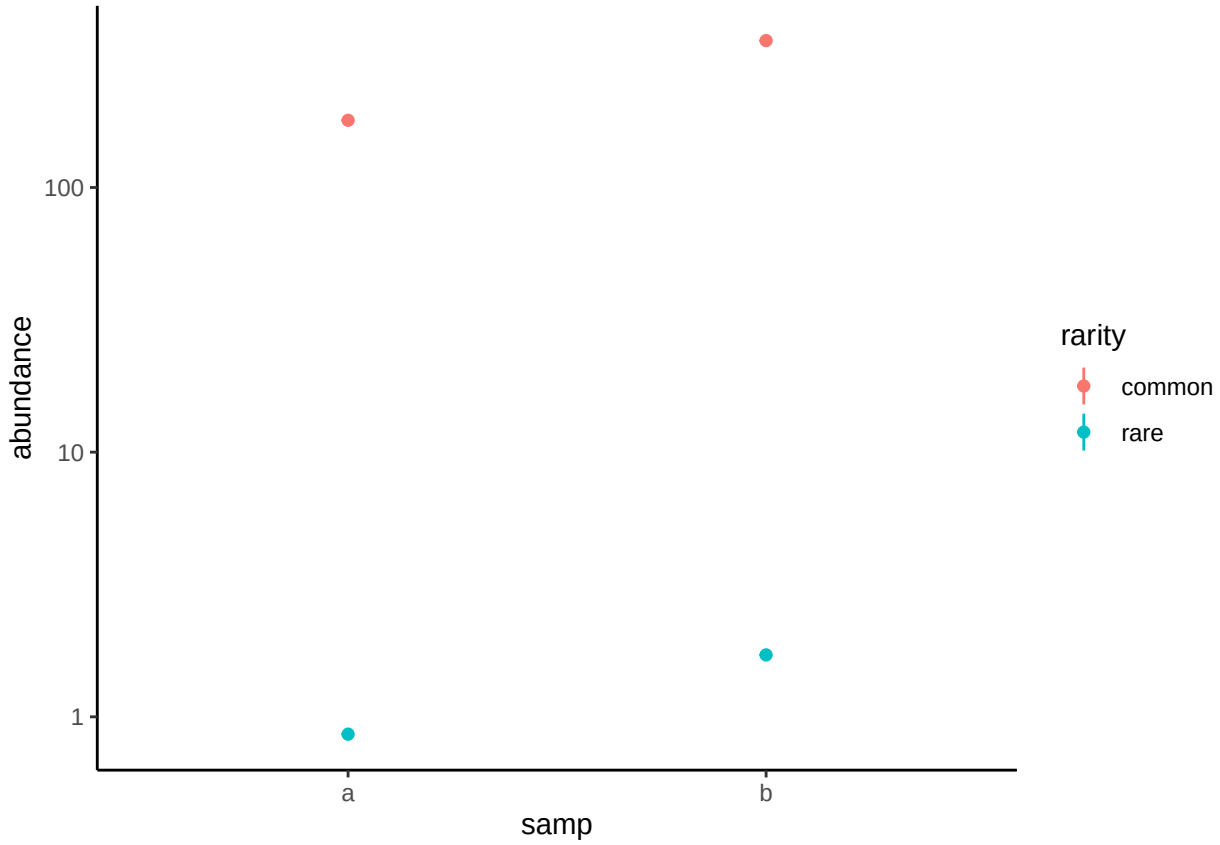
```
set.seed(666) # guarantees demonic outcomes
reps <- 9999 # set high so we're not caught off guard by noise
a <- rmultinom(reps, 200, prob = regional_abundances)
b <- rmultinom(reps, 400, prob = regional_abundances)
a_common<- a[1:100,]
a_rare <- a[301:400,]
b_common <- b[1:100,]
b_rare <- b[301:400,]
```

I expect to see double the number of individuals from rare species in "b" as "a", and also double the number of individuals from common species. Is this what you expect?

```
abundance_data <- map_dfr(c("a", "b"), function(samp){
  map_dfr(c("rare", "common"), function(rarity){
    data.frame(abundance = apply(get(paste(samp, rarity, sep = "_"))
                , 2, sum)
    , samp = samp
```

```
    , rarity = rarity
    )
  })
})

# look at means and se
abundance_data %>%
  ggplot(aes(samp, abundance, color = rarity)) +
  stat_summary(fun = mean
               , fun.max = function(x){mean(x) +  sd(x)/sqrt(length(x))}
               , fun.min = function(x){max(mean(x) - sd(x)/sqrt(length(x)), 0)}
               , size = 0.25) +
  theme_classic() +
  coord_trans(y = "log10") +
  scale_y_continuous(breaks = scales::breaks_log())
```



Yes, In this picture it looks like the rare species increase as much as the common ones did. What does a model tell us about the average increase?

```
ab_mod <- MASS::glm.nb(abundance ~ rarity*samp
               , data = abundance_data
               , control = glm.control(maxit = 999)
               )
```

```
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
## control$trace > : iteration limit reached
```

```
## Warning in sqrt(1/i): NaNs produced

## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
## control$trace > : iteration limit reached

## Warning in sqrt(1/i): NaNs produced
```

```
ab_mod$theta # use poisson
```

```
## [1] 2.120561e+17
```

```
ab_mod <- glm(abundance ~ rarity*samp
              , data = abundance_data
              , family = "poisson"
              )
```

If things went as I would expect, I'd see a significant effect of `samp` (approximate doubling) and a significant effect of `rarity` – and no interaction. Is that what I see?

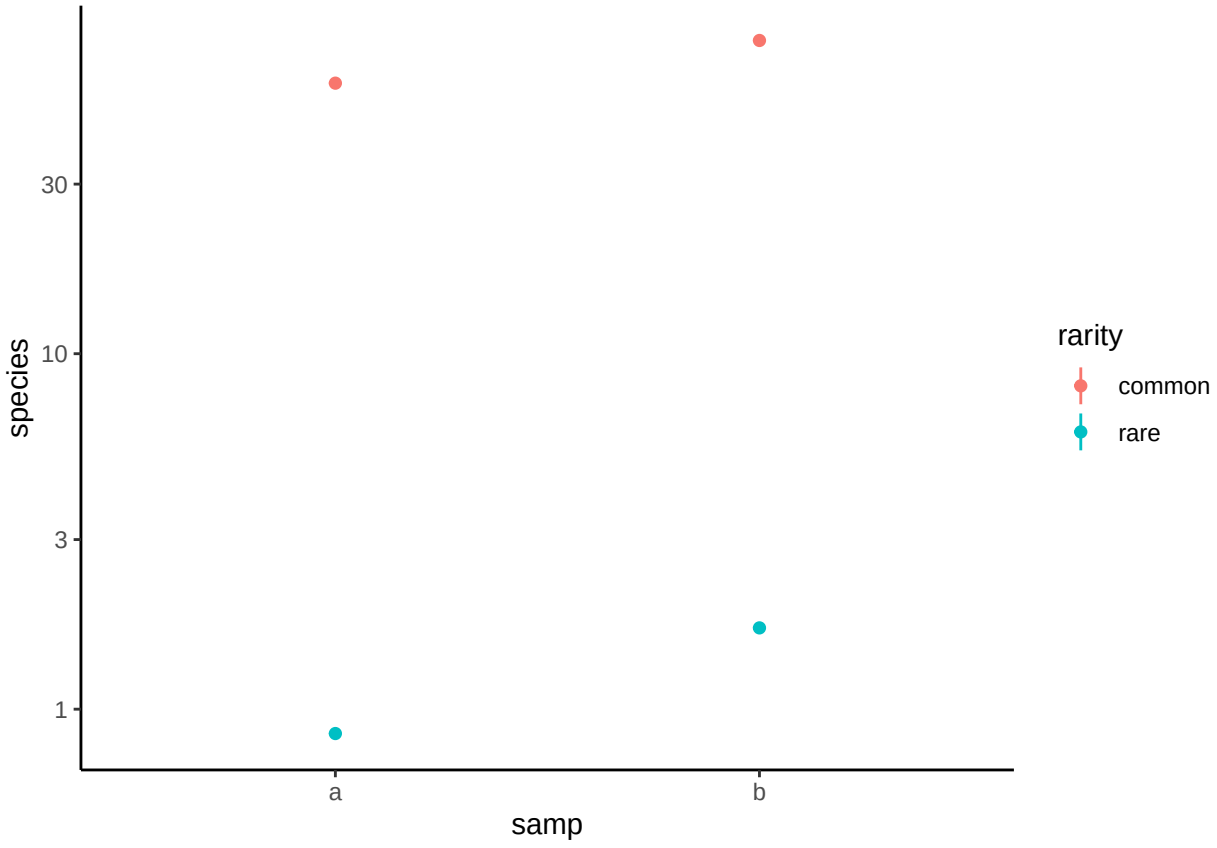
```
summary(ab_mod)
```

```
##
## Call:
## glm(formula = abundance ~ rarity * samp, family = "poisson",
##     data = abundance_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8513  -0.5598   0.0412   0.2130   3.9092
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        5.1898807  0.0007465 6951.91   <2e-16 ***
## rarityrare        -5.3403711  0.0108078 -494.12   <2e-16 ***
## sampb              0.6932305  0.0009143  758.20   <2e-16 ***
## rarityrare:sampb -0.0041021  0.0132456   -0.31    0.757
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 7809343  on 39995  degrees of freedom
## Residual deviance:   24874  on 39992  degrees of freedom
## AIC: 206255
##
## Number of Fisher Scoring iterations: 5
```

It is what I see! Note that the interaction term is both small and not significant; this means we're seeing basically the same change of about 2x in abundance of "rare" and "common" taxa. Unfortunately, my expectations might not carry over to the change in the number of *species* I see between `a` and `b`.

```r
ab_r_data <- map_dfr(c("a", "b"), function(samp){
  map_dfr(c("rare", "common"), function(rarity){
    data.frame(abundance = apply(get(paste(samp, rarity, sep = "_"))
                , 2, sum)
               , species = apply(get(paste(samp, rarity, sep = "_"))
                , 2, function(x){sum(x>0)})
    , samp = samp
    , rarity = rarity
    )
  })
})

ab_r_data %>%
  ggplot(aes(samp, species, color = rarity)) +
  stat_summary(fun = mean
               , fun.max = function(x){mean(x) +  sd(x)/sqrt(length(x))}
               , fun.min = function(x){max(mean(x) - sd(x)/sqrt(length(x)), 0)}
               , size = 0.25) +
  theme_classic() +
  coord_trans(y = "log10") +
  scale_y_continuous(breaks = scales::breaks_log())
```



Ok, now this is bugging me. It looks like I see a **bigger** increase in the number of rare species between `a` and `b` than I do in the number of common species. But of course, nothing about the difference in sampling intensity applied to rare species but not common ones, right?

Is this what a model says, too?

```
# treating the count of species as... a count
# species_mod <- MASS::glm.nb(species ~ rarity*samp
#               , data = ab_r_data
#               , control = glm.control(maxit = 999)
#               )
#
#
# species_mod$theta # use poisson

species_mod <- glm(species ~ rarity*samp
              , data = ab_r_data
              , family = "poisson"
              )

summary(species_mod)
```

```
##
## Call:
## glm(formula = species ~ rarity * samp, family = "poisson", data = ab_r_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0261  -0.5772   0.0410   0.3023   3.9316
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       4.055057   0.001317 3079.78   <2e-16 ***
## rarityrare       -4.212898   0.010901 -386.45   <2e-16 ***
## sampb             0.276626   0.001746  158.44   <2e-16 ***
## rarityrare:sampb  0.407818   0.013387   30.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1691186  on 39995  degrees of freedom
## Residual deviance:   26775  on 39992  degrees of freedom
## AIC: 181204
##
## Number of Fisher Scoring iterations: 5
```

Yikes! Now the interaction is huge! I mean, it's not wrong... You probably saw this coming... a large fraction of the common species were present in `a`, so that fraction (and thus the counts) just can't increase that much in `b`. By contrast, only a percent or two of the rare species is typically found in `a`, so there is a lot of room to grow!

If we think about an increase in the *fraction* of the total species found in `a` vs. `b`, maybe we should be thinking of a logistic regression instead.

```
ab_r_data <- ab_r_data %>%
  mutate(r_perc = species/100)
species_logistic <- glm(r_perc~samp*rarity
                        , data = ab_r_data
```

```
                        , family = "binomial"
                        , weights = rep.int(100, length(abundance_data[,1])))
summary(species_logistic)
```

```
##
## Call:
## glm(formula = r_perc ~ samp * rarity, family = "binomial", data = ab_r_data,
##     weights = rep.int(100, length(abundance_data[, 1])))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3415  -0.7102   0.0631   0.4691   4.0018
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.309998   0.002024  153.15   <2e-16 ***
## sampb             0.846645   0.003097  273.37   <2e-16 ***
## rarityrare       -5.064432   0.011055 -458.11   <2e-16 ***
## sampb:rarityrare -0.153701   0.013704  -11.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2435062  on 39995  degrees of freedom
## Residual deviance:   36040  on 39992  degrees of freedom
## AIC: 167228
##
## Number of Fisher Scoring iterations: 5
```

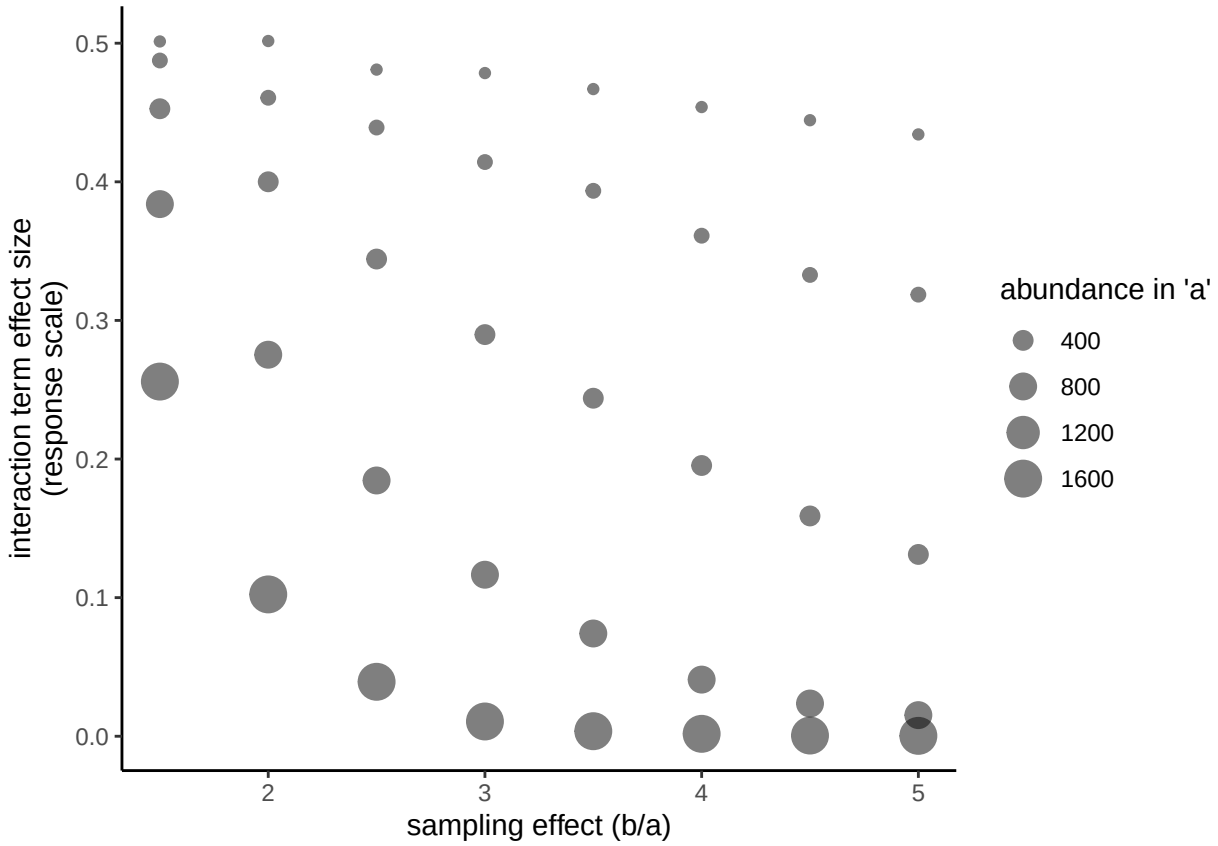
Ok, In an earlier moment, this looked how I wanted, and now it doesn't! I'm worried I kinda got lucky there... does the lack of interaction depend on the size of the sampling effect? On the sampling intensities in absolute terms?

```
samp_eff <- seq(1.5, 5, 0.5)
base_ab <- c(100, 200, 400, 800, 1600)
reps <- 9999
eff_trend <- map_dfr(base_ab, function(ba){
  map_dfr(samp_eff, function(seff){
  a <- rmultinom(reps, ba, prob = regional_abundances)
  b <- rmultinom(reps, ba
                 *seff, prob = regional_abundances)
  a_common <- a[1:100,]
  a_rare <- a[301:400,]
  b_common <- b[1:100,]
  b_rare <- b[301:400,]
  dd <- map_dfr(c("a", "b"), function(samp){
  map_dfr(c("rare", "common"), function(rarity){
    data.frame(abundance = apply(get(paste(samp, rarity, sep = "_"))
                , 2, sum)
              , species = apply(get(paste(samp, rarity, sep = "_"))
                , 2, function(x){sum(x>0)})
              , samp = samp
```

```
              , rarity = rarity
              , base_abund = ba
              , sampling_effect = seff
    ) %>% mutate(rperc = species/100)
      })
    })
  })
})
eff_mods <- eff_trend %>%
  mutate(wt = 100) %>%
  nest_by(base_abund, sampling_effect) %>%
  mutate(my_mod = list(glm(rperc~samp*rarity
                           , data = data
                           , family = "binomial"
                           , weights = wt)))
effs <- eff_mods %>% summarize(intx = summary(my_mod)$coefficients[4,1]
            , pval = summary(my_mod)$coefficients[4,4])
```

```
## `summarise()` has grouped output by 'base_abund', 'sampling_effect'. You can
## override using the `.groups` argument.
```

```
effs %>%
  ggplot(aes(sampling_effect
             , boot::inv.logit(intx)
             , size =  base_abund)) +
  geom_point(alpha = 0.5) +
  theme_classic() +
  labs(x = "sampling effect (b/a)"
       , y = "interaction term effect size\n(response scale)"
       , size = "abundance in 'a'") +
  scale_color_viridis_c() +
  scale_size_area()
```



FUDGE! it totally does. Let's look at the graph:

- Interaction shrinks with the size of the sampling effect
- Interaction shrinks with sample sizes

One thought: Why am I not using a beta-binomial?

```
# not sure this is working, skip
# beta_mods <- eff_trend %>%
#   mutate(wt = 100) %>%
#   nest_by(base_abund, sampling_effect) %>%
#   mutate(my_mod = list(glmmTMB::glmmTMB(rperc~samp*rarity
#                          , data = data
#                          , family = glmmTMB::betabinomial
#                          , weights = wt)))
#
# betas <- beta_mods %>% summarize(intx = summary(my_mod)$coefficients$cond[4,1]
#              , pval = summary(my_mod)$coefficients$cond[4,4])
# betas %>%
#   ggplot(aes(sampling_effect
#              , pval
#              , color = boot::inv.logit(intx)
#              , size =  base_abund)) +
```

```
#   geom_point(alpha = 0.5) +
#   theme_classic() +
#   labs(x = "sampling effect (b/a)"
#        , y = "p value for sampling  x rarity coefficient"
#        , color = "interaction term effect size\n(response scale)"
#        , size = "abundance in 'a'") +
#   scale_color_viridis_c() +
#   scale_size_area()
```

I'm a bit stuck on how to proceed

I don't want to pick up on a significant interaction between the sampling intensity (which is real) and the rarity (which is also real) when, in my mind, they shouldn't be interacting! But I don't have a clear idea in my head about what the appropriate model would be for looking at changes in the number of observed species when sampling intensity is larger in one treatment than another.

Sandbox.... Why would't you just sample the same in all treatments?

I sample bees by looking into flowers in timed, fixed-area surveys, and if there is a bee contacting the reproductive parts of the flower, I capture the bee. From one perspective, the amount of sampling I do is about the time and area dimensions. But the number of bees I sample will increase if I look into more flowers in a given time and area... assuming the visit rate per flower stays pretty constant. I think of that increase as a sampling effect, even if it is mediated by a treatment that increases the flowers per area.

On the other hand, I'd be very interested if planting particular kinds of flowers attracted some kinds of bees that would otherwise not come by the yard.