Discovering Associations

Zane Kliesmete

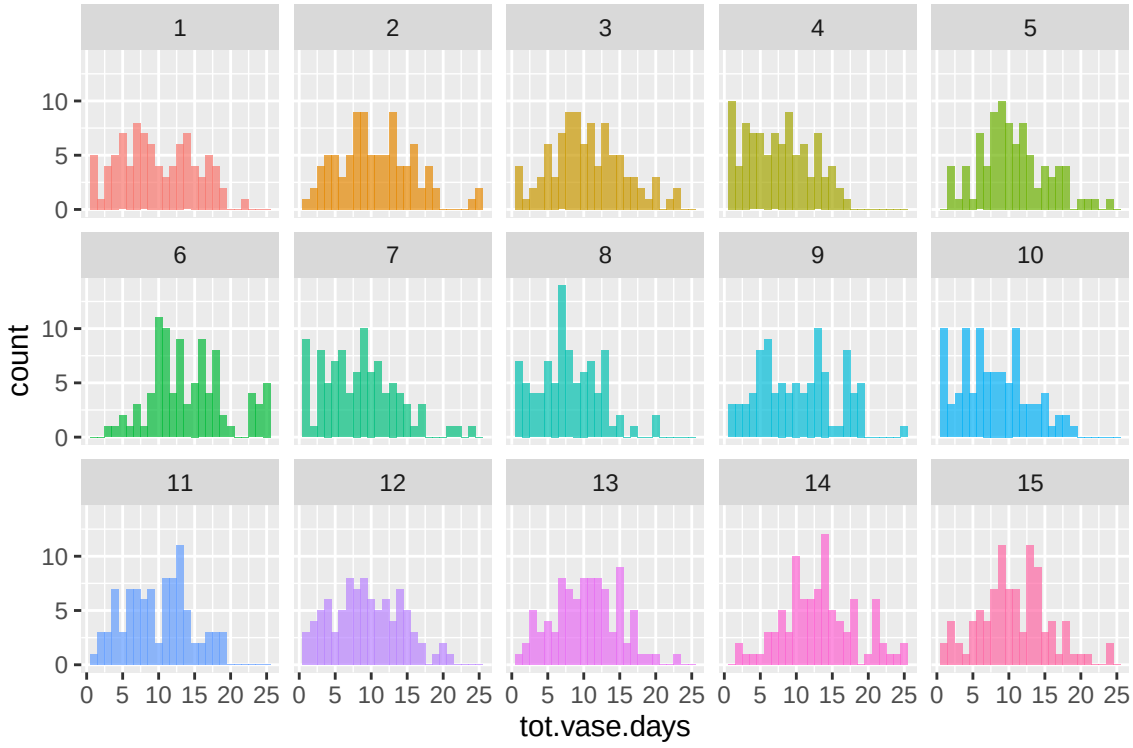
Count data of rose vase life days.

Poisson-mixed effects modelling.

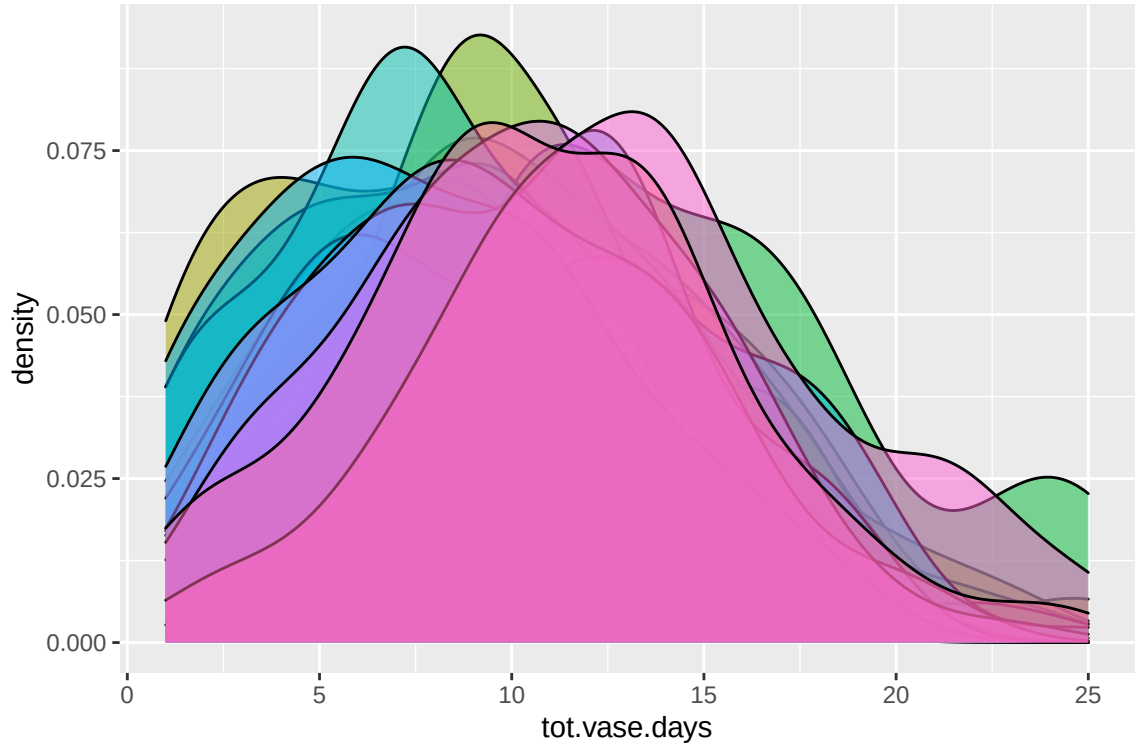
Out of 1440 datapoints, we have 60 missing outcomes (total vase days), which is 4.1% of the data. More data description, can copy some from simulation description.

Checking the distribution of the Vase Days of the count data. It looks like there is a larger degree of overdispersion than the poisson model currently accounts for (nope, it's just conditional on the compound: the goodfit should be done per compound).

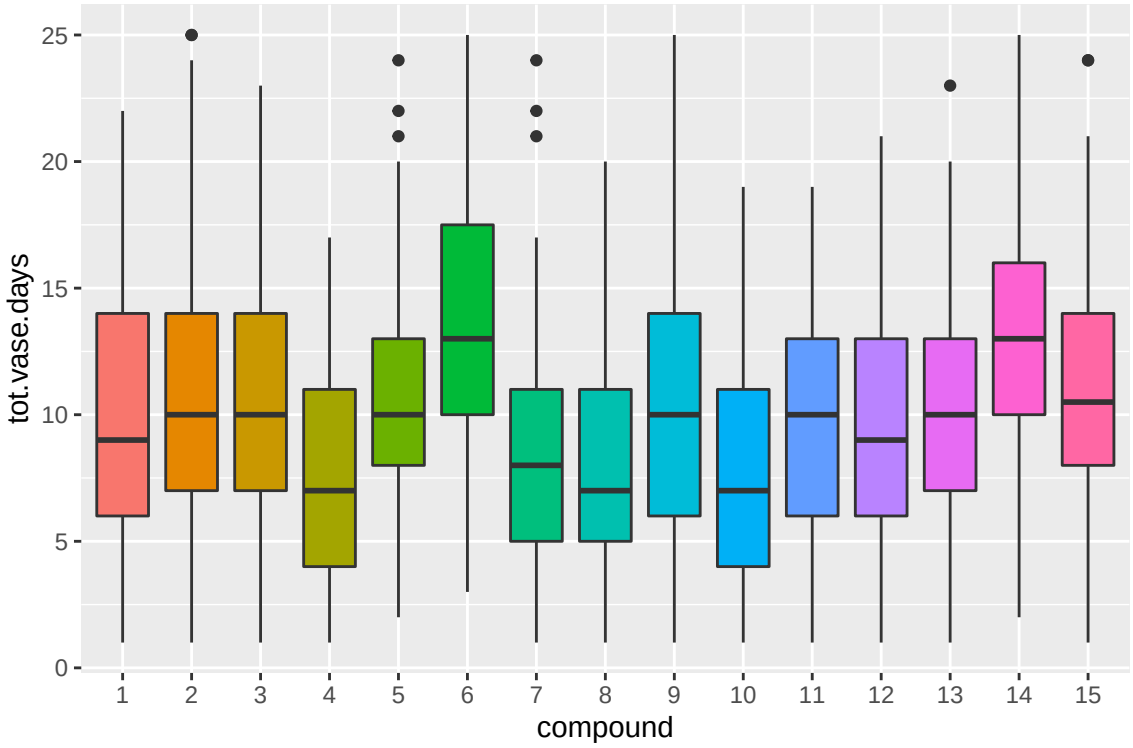
```
ggplot(d, aes(x=tot.vase.days, fill=compound))+geom_histogram(alpha=0.7, bins=25)+facet_wrap(~compound,
```




```
ggplot(d, aes(x=tot.vase.days, fill=compound))+geom_density(alpha=0.5)+theme(legend.position = "none")
```



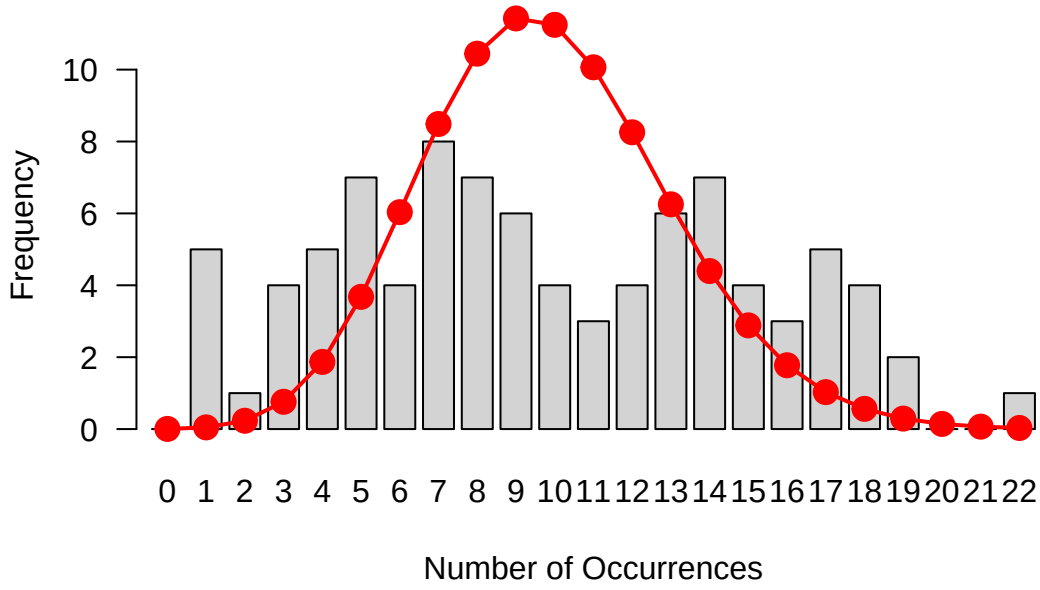

```
#ggplot(d %>% filter(compound %in% c(1, 6, 14)), aes(x=tot.vase.days, fill=as.factor(compound)))+geom_d

# d %>% group_by(compound) %>%
#   dplyr::summarise(mean=mean(tot.vase.days),
#                    var=var(tot.vase.days)) %>%
#   ggplot(aes(x=mean, y=var))+
#   geom_abline(slope=1, intercept=0)+
#   geom_point()
ggplot(d, aes(x=compound,y=tot.vase.days, fill=compound))+geom_boxplot()+theme(legend.position = "none"
```




```
gf <- goodfit(d$tot.vase.days[d$compound==1], "poisson")
summary(gf)
```

```
##
##   Goodness-of-fit test for poisson distribution
##
##                       X^2 df     P(> X^2)
## Likelihood Ratio 100.8575 18 1.542306e-13
```

```
plot(gf, type="standing", scale="raw") #I guess the fit would be better if we did it per compound
```




```
#https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#overdispersion
overdisp_fun <- function(model) {
rdf <- df.residual(model)
rp <- residuals(model,type="pearson")
Pearson.chisq <- sum(rp^2)
prat <- Pearson.chisq/rdf #pearson residuals over the residual degrees of freedom
pval <- pchisq(Pearson.chisq, df=rdf, lower.tail=FALSE)
c(chisq=Pearson.chisq,ratio=prat,rdf=rdf,p=pval)
}


#inspired by the first answer here for nested data formulation https://stats.stackexchange.com/question
glmer_out <- glmer(tot.vase.days ~ compound + species + garden + (1|rater) + (1|subplotID/bushID), fami
overdisp_fun(glmer_out) #quite oki! no significance, yey. i think this might be more relevant than good
```

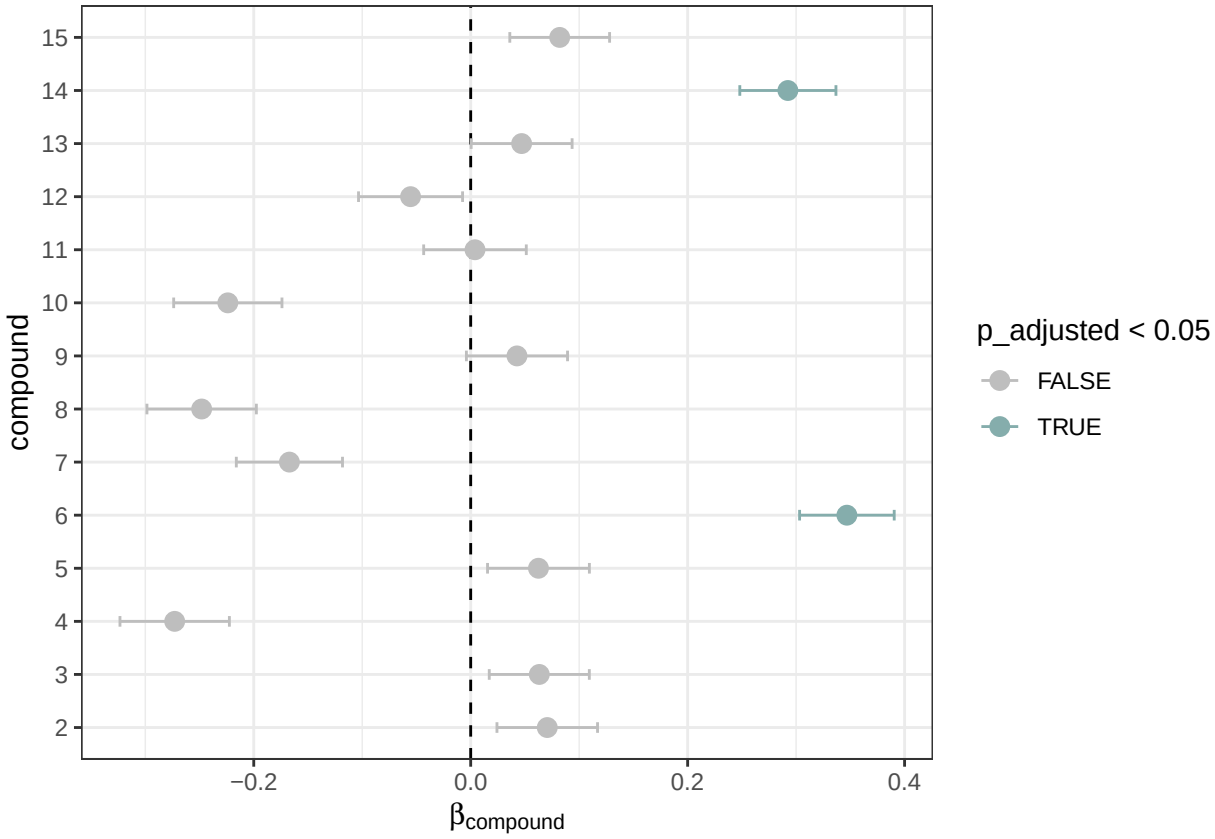
```
##        chisq        ratio          rdf            p
## 1416.6181038    1.0416310 1360.0000000    0.1392638
```

```
#https://fukamilab.github.io/BIO202/04-B-binary-data.html#glm_for_count_data
drop1(glmer_out, test = "Chi")
```

```
## Single term deletions
##
## Model:
## tot.vase.days ~ compound + species + garden + (1 | rater) + (1 |
##     subplotID/bushID)
##          npar    AIC    LRT Pr(Chi)
## <none>        7222.6
```

```
## compound   14 7615.7 421.07  <2e-16 ***
## species     1 7222.1   1.43  0.2315
## garden      1 7222.9   2.23  0.1356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
glmer_coefficients<-as.data.frame(summary(glmer_out)$coeff) %>%
  rownames_to_column("predictor") %>%
  filter(grepl("compound",predictor)) %>%
  dplyr::rename(pval=`Pr(>|z|)`) %>%
  #we want to have p-adjusted (Holm) values for one-sided test H.alt: lambda(compound)>lambda(water)
  dplyr::mutate(one_sided_pval=ifelse(`z value`>0, pval/2, (1-pval/2)),
         p_adjusted=p.adjust(one_sided_pval, method="holm"),
         significant_higher=ifelse(p_adjusted<0.05, T, F))

ggplot(glmer_coefficients %>%
         mutate(compound=factor(gsub("compound","",predictor), levels=2:15)),
       aes(x=compound, y=Estimate, color=p_adjusted<0.05))+
  geom_hline(yintercept=0, linetype="dashed")+
  geom_errorbar(aes(ymin=Estimate - `Std. Error`, ymax=Estimate +`Std. Error`), width=0.2)+geom_point(s
  scale_color_manual(values=c("grey","#85ADAC"))+coord_flip()+
  ylab(expression(beta["compound"]))
```



Conclusion: compounds 6 and 14 significantly increase rose vase days (mention estimates +- sd error, also backcalculated in days, alpha, one-sided Wald test, maybe the exact z and p-values, correction Holm).

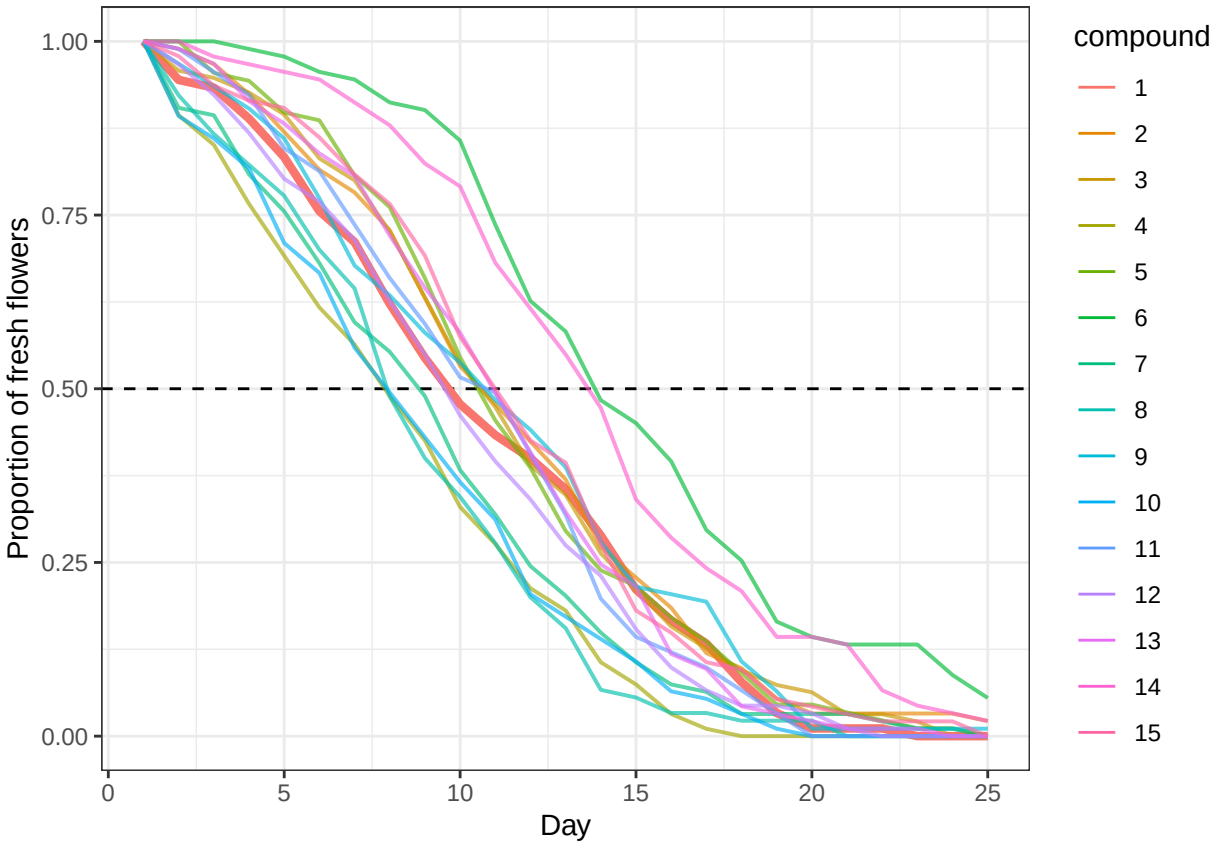
Binomial longitudinal modelling.

Fit a longitudinal binary data predicting vase life. First need to transform the data into a binary outcome per day.

```
outmat<-matrix(nrow = nrow(d), ncol=max(d$tot.vase.days))

outmat[is.na(outmat)]<-1
for (i in 1:nrow(outmat)){
  outmat[i,c(d[i,tot.vase.days]:25)]<-0
  outmat[i,d[i,tot.vase.days]]<-1

}

outdf<-as.data.frame(outmat)
names(outdf)<-paste0("newVar_",names(outdf))

d_full<-d %>%
  bind_cols(outdf %>% as.data.frame()) %>%
  pivot_longer(contains("newVar"), names_to="day", values_to = "fresh") %>%
  mutate(day=as.numeric(gsub("newVar_V","",day)))

data_full_cc <- aggregate(fresh ~ compound + day, data = d_full, FUN = mean) %>%
  mutate(water=ifelse(compound==1,T,F))

ggplot(data = data_full_cc)+
  geom_hline(yintercept=0.5, linetype="dashed")+
    geom_line(aes(x = day, y = fresh, color = compound, size=water, alpha=water)) +
  scale_size_discrete(range=c(0.7,1.5),guide="none")+
  scale_alpha_discrete(range=c(0.65,1), guide="none")+
  theme_bw()+
  ylab("Proportion of fresh flowers")+
  xlab("Day")
```

```
## Warning: Using size for a discrete variable is not advised.
```

```
## Warning: Using alpha for a discrete variable is not advised.
```




```
#included this just to see if all datapoints are there..
#ggplot(data = data_full_cc)+geom_line(aes(x = day, y = fresh, color = compound))+facet_wrap(~compound)
```

```
#xtabs(~ garden + subplotID, d_full)

# we add the day as a random effect, because the likelihood to "switch" from
glmer_out_bn <- glmer(fresh ~ compound + species + garden + (1|day) + (1|rater) + (1|subplotID/bushID),
summary(glmer_out_bn)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: fresh ~ compound + species + garden + (1 | day) + (1 | rater) +
##     (1 | subplotID/bushID)
##    Data: d_full
##
##      AIC      BIC   logLik deviance df.resid
##  15667.5  15844.9  -7812.8  15625.5    34479
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -14.6291  -0.1854  -0.0284   0.1218  12.0753
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
```

```
##  bushID:subplotID (Intercept)  0.3137  0.560
##  day              (Intercept) 20.6604  4.545
##  subplotID        (Intercept)  1.5869  1.260
##  rater            (Intercept)  3.7533  1.937
## Number of obs: 34500, groups:
## bushID:subplotID, 96; day, 25; subplotID, 16; rater, 6
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.41750    1.28191  -1.106 0.268825
## compound2    0.42821    0.11331   3.779 0.000157 ***
## compound3    0.38181    0.11248   3.394 0.000688 ***
## compound4   -1.47691    0.11809 -12.506  < 2e-16 ***
## compound5    0.36618    0.11453   3.197 0.001387 **
## compound6    2.29454    0.11419  20.094  < 2e-16 ***
## compound7   -0.93361    0.11585  -8.058 7.73e-16 ***
## compound8   -1.36647    0.11869 -11.512  < 2e-16 ***
## compound9    0.25937    0.11331   2.289 0.022078 *
## compound10  -1.22496    0.11731 -10.442  < 2e-16 ***
## compound11   0.02115    0.11427   0.185 0.853144
## compound12  -0.31993    0.11501  -2.782 0.005409 **
## compound13   0.27980    0.11324   2.471 0.013475 *
## compound14   1.88567    0.11369  16.586  < 2e-16 ***
## compound15   0.49578    0.11256   4.405 1.06e-05 ***
## species2    -0.16720    0.12190  -1.372 0.170183
## garden2      0.97666    0.64120   1.523 0.127713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Correlation matrix not shown by default, as p = 17 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

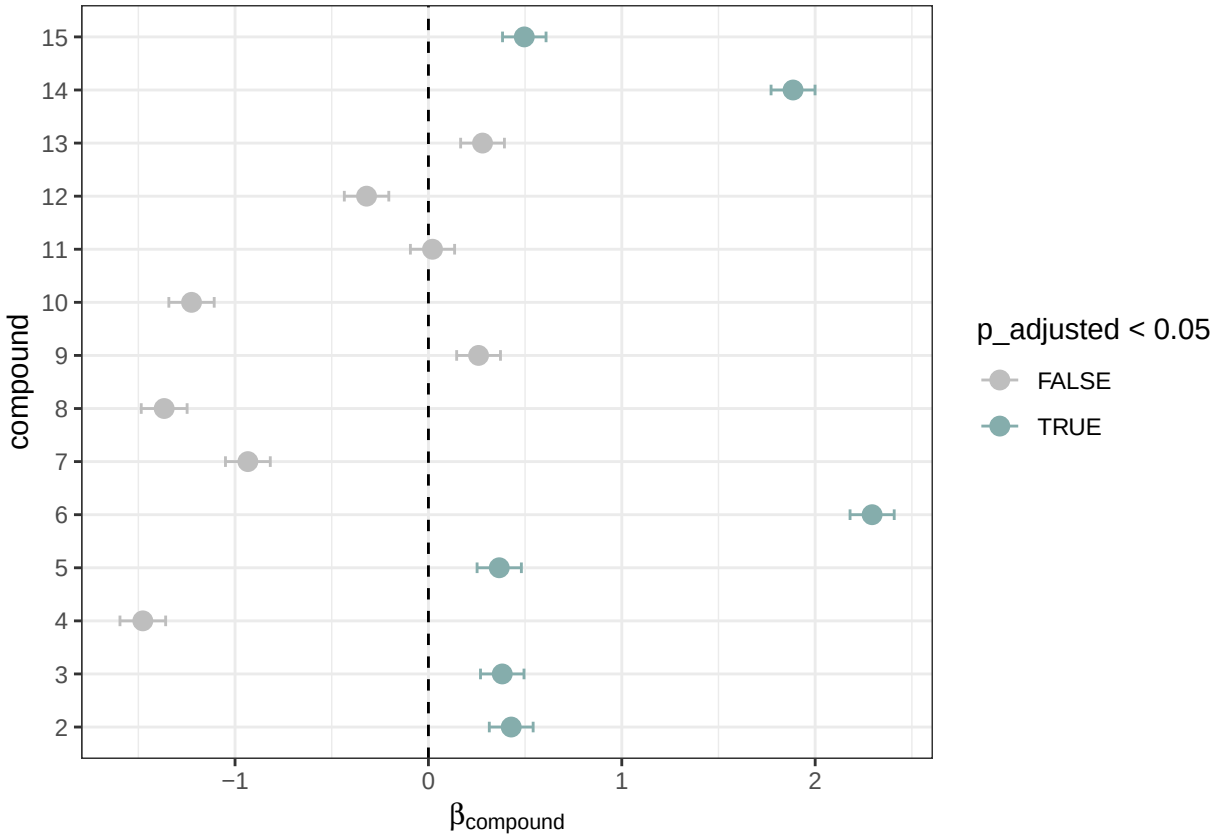
```
#drop1(glmer_out_bn) this does not converge when dropping species


#this model also works, so i guess just putting lowerID AND the day into the same model is an overkill,
glmer_out_bn2 <- glmer(fresh ~ compound + species + garden + (1|flowerID) + (1|rater) + (1|subplotID/bus
summary(glmer_out_bn2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: fresh ~ compound + species + garden + (1 | flowerID) + (1 | rater) +
##     (1 | subplotID/bushID)
##    Data: d_full
##
##      AIC      BIC   logLik deviance df.resid
##  42243.4  42420.8 -21100.7  42201.4    34479
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.4998 -0.7717 -0.4650  0.9090  4.1585
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  flowerID         (Intercept) 0.1896   0.4354
##  bushID:subplotID (Intercept) 0.0128   0.1131
##  subplotID        (Intercept) 0.1709   0.4134
##  rater            (Intercept) 0.3681   0.6067
## Number of obs: 34500, groups:
## flowerID, 1380; bushID:subplotID, 96; subplotID, 16; rater, 6
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.63774    0.29538  -2.159  0.03085 *
## compound2    0.15060    0.09216   1.634  0.10222
## compound3    0.12966    0.09144   1.418  0.15619
## compound4   -0.47791    0.09353  -5.110 3.22e-07 ***
## compound5    0.12872    0.09297   1.385  0.16618
## compound6    0.79342    0.09233   8.594  < 2e-16 ***
## compound7   -0.30417    0.09296  -3.272  0.00107 **
## compound8   -0.43403    0.09416  -4.609 4.04e-06 ***
## compound9    0.08306    0.09215   0.901  0.36739
## compound10  -0.40144    0.09350  -4.293 1.76e-05 ***
## compound11   0.01684    0.09245   0.182  0.85550
## compound12  -0.10101    0.09288  -1.088  0.27680
## compound13   0.10041    0.09180   1.094  0.27406
## compound14   0.64369    0.09222   6.980 2.96e-12 ***
## compound15   0.17098    0.09146   1.869  0.06157 .
## species2    -0.04555    0.04086  -1.115  0.26492
## garden2      0.31811    0.21034   1.512  0.13045
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Correlation matrix not shown by default, as p = 17 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

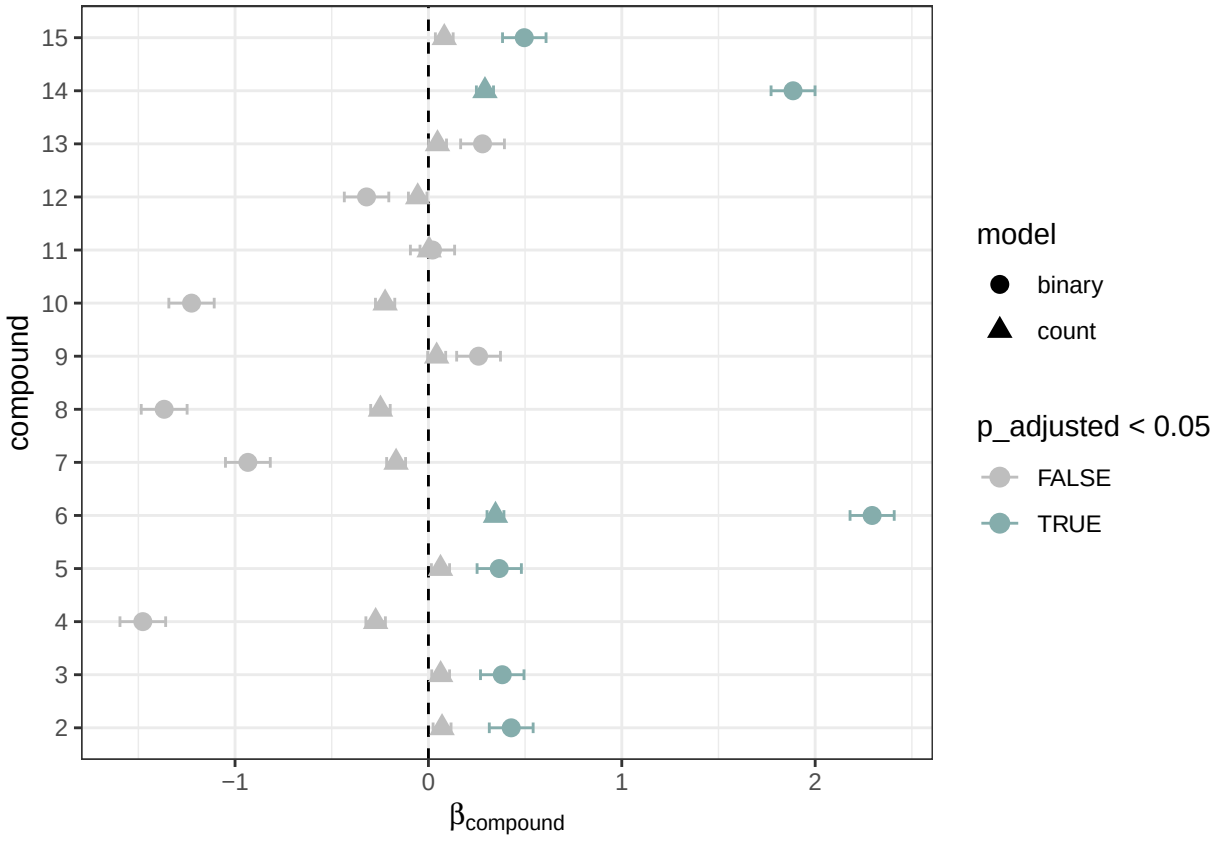
```
#this does not converge
#glmer_out_bn_flowerID <- glmer(fresh ~ compound + species + garden + (1|flowerID) + (1|day) + (1|rater
#summary(glmer_out_bn_flowerID)


#glmer_out_bn_bild <- glmer(fresh ~ compound+day + species + (1|rater) + (1|garden/subplotID), family=b


glmer_bn_coefficients<-as.data.frame(summary(glmer_out_bn)$coeff) %>%
  rownames_to_column("predictor") %>%
  filter(grepl("compound",predictor)) %>%
  #filter(grepl("day",predictor)) %>%
  dplyr::rename(pval=`Pr(>|z|)`) %>%
  #we want to have p-adjusted (Holm) values for one-sided test
  #however in this case it's the moe the day decreases, the more the probability of 1 should increase--
  #H.alt: lambda(compound)<lambda(water)
```

```
  dplyr::mutate(one_sided_pval=ifelse(`z value`>0, pval/2, (1-pval/2)),
         p_adjusted=p.adjust(one_sided_pval, method="holm"),
         significant_higher=ifelse(p_adjusted<0.05, T, F))

ggplot(glmer_bn_coefficients %>%
         mutate(compound=factor(gsub("compound","",predictor), levels=2:15)),
       aes(x=compound, y=Estimate, color=p_adjusted<0.05))+
  geom_hline(yintercept=0, linetype="dashed")+
  geom_errorbar(aes(ymin=Estimate - `Std. Error`, ymax=Estimate +`Std. Error`), width=0.2)+geom_point(si
  scale_color_manual(values=c("grey","#85ADAC"))+coord_flip()+
  ylab(expression(beta["compound"]))
```




```
both_predictions1<-bind_rows(glmer_coefficients %>% mutate(model="count"),
                             glmer_bn_coefficients %>% mutate(model="binary"))


#both_predictions1<-inner_join(glmer_coefficients, glmer_bn_coefficients, by="predictor", suffix=c(".co
  #mutate(significant_in_either=ifelse(significant_higher | significant, T, F))

ggplot(both_predictions1 %>%
         mutate(compound=factor(gsub("compound","",predictor), levels=2:15)),
       aes(x=compound, y=Estimate, color=p_adjusted<0.05, shape=model))+
  geom_hline(yintercept=0, linetype="dashed")+
  geom_errorbar(aes(ymin=Estimate - `Std. Error`, ymax=Estimate +`Std. Error`), width=0.2)+geom_point(si
  scale_color_manual(values=c("grey","#85ADAC"))+coord_flip()+
```

```
  ylab(expression(beta["compound"]))
```



Gaussian data of flower width.

Gaussian outcome data. We received data from 180 flowers. This was distributed as 12 flowers for each compound for each of the 15 compounds. In each of those groups, there were 6 flowers per species and 6 grown in each garden. There were also 18 different subplots. The number of subplots is greater than the number of number of flowers per group.

For each of the 18 flowers, we have measurements of the width of the flower over the course of 21 days. All measurements for all flowers were taken by a single rater.

Below I transform the data so that there is a row for each measurement of each flower on each day resulting in 3780 rows.

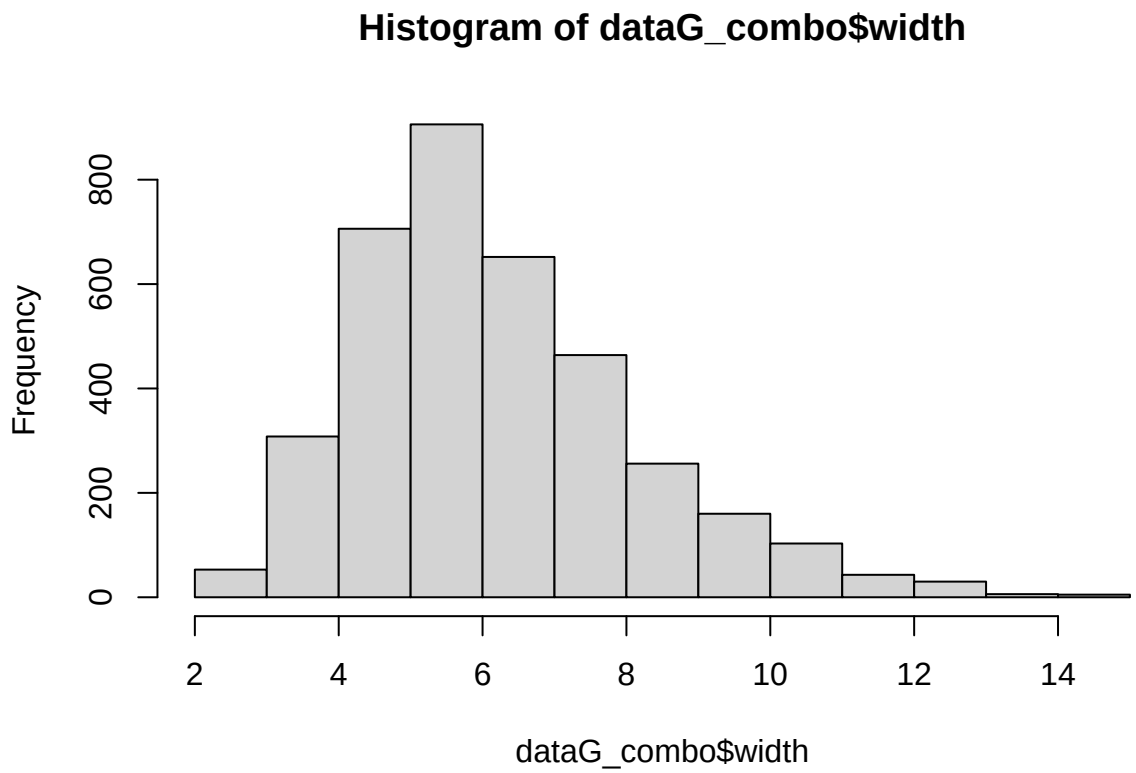
```
g <- fread('gaussian_data_G6.csv')
#summary(g) #there is only one rater, drop it
g<-g %>% dplyr::select(-Rater)


colnames(g)<-c("flowerID",0:20,"compound","type","garden","subplot")

dataG_long <- gather(g, days, width, "0":"20", factor_key=TRUE) %>%
      mutate(garden=as.factor(garden),
             type=as.factor(type),
```

```
            compound=as.factor(compound),
            subplot=as.factor(subplot),
            days=as.numeric(days),
            flowerID=as.factor(flowerID))

head(dataG_long)
```

```
##   flowerID compound type garden subplot days width
## 1    18075        1    1      1       1    1   2.9
## 2    18767        1    1      1       2    1   2.6
## 3    18028        1    1      1       3    1   5.2
## 4    18326        1    1      2       4    1   6.5
## 5    18017        1    1      2       5    1   4.2
## 6    18718        1    1      2       6    1   5.7
```

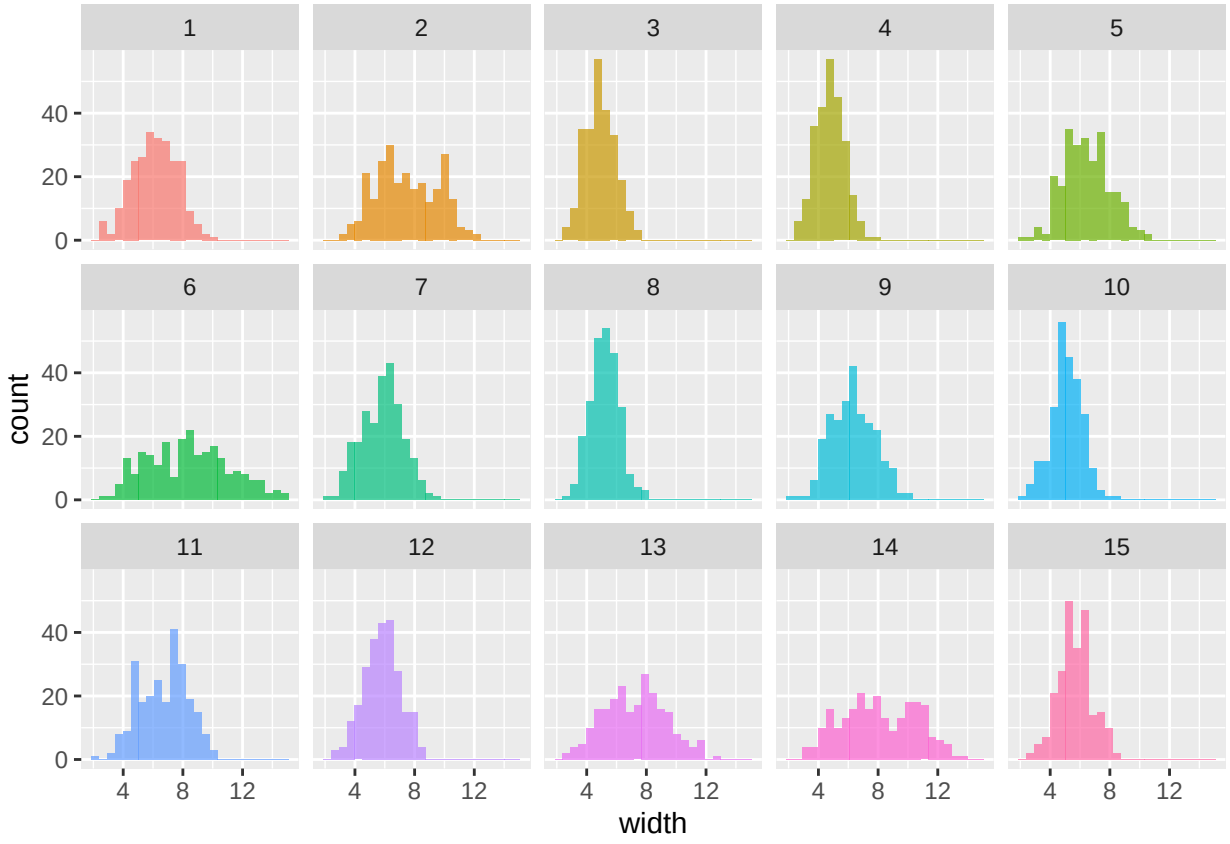
I also added a column showing the change in the width of the flower so that we can see the change in width per day. It is worth noting that the width of the flower does not uniformly increase, instead it does fluctuate from day to day, decreasing occasionally. Also, there are quite a few missing measurements, we probably should have accounted for this in our sample size calculation?

```
hist(dataG_combo$width)
```



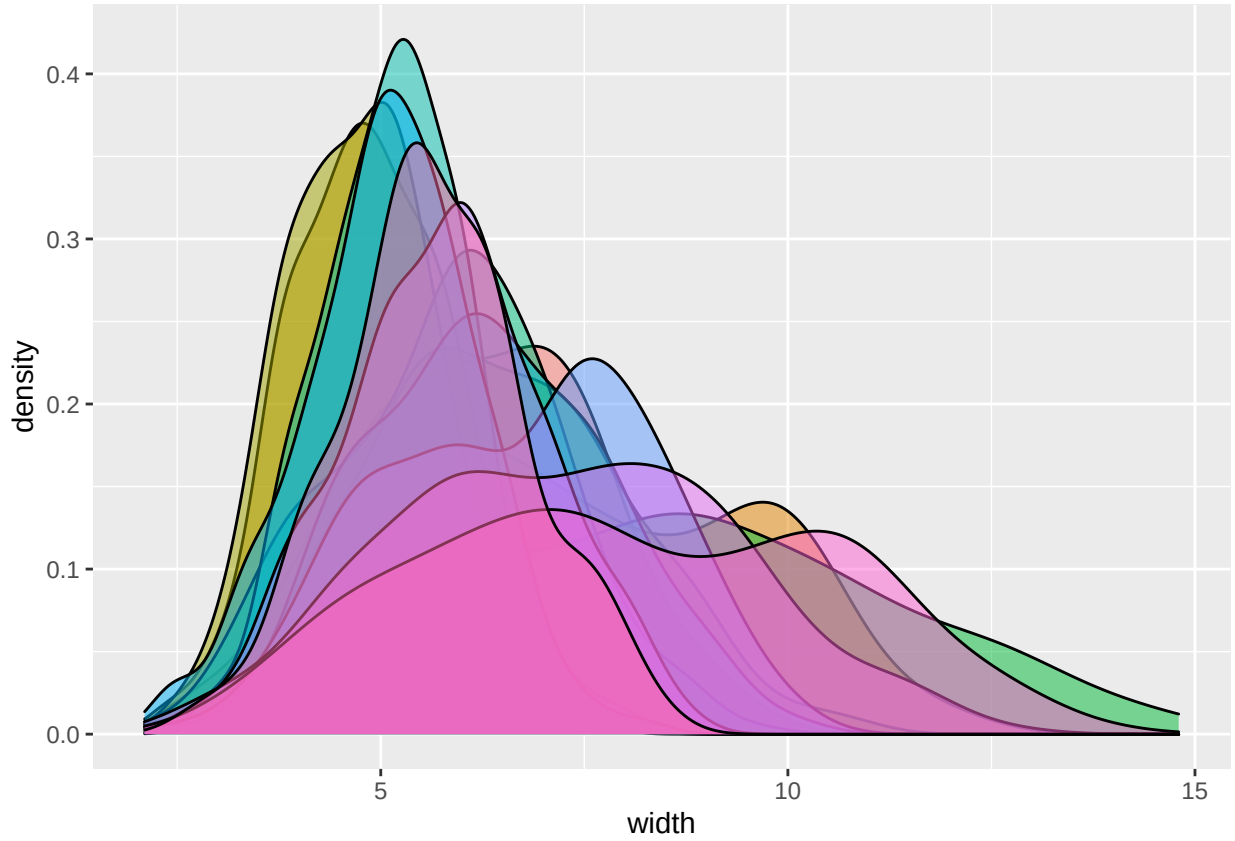

```
#hist(dataG_combo$delta_width)

ggplot(dataG_combo, aes(x=width, fill=compound))+geom_histogram(alpha=0.7, bins=25)+facet_wrap(~compound
```

```
## Warning: Removed 88 rows containing non-finite values (stat_bin).
```



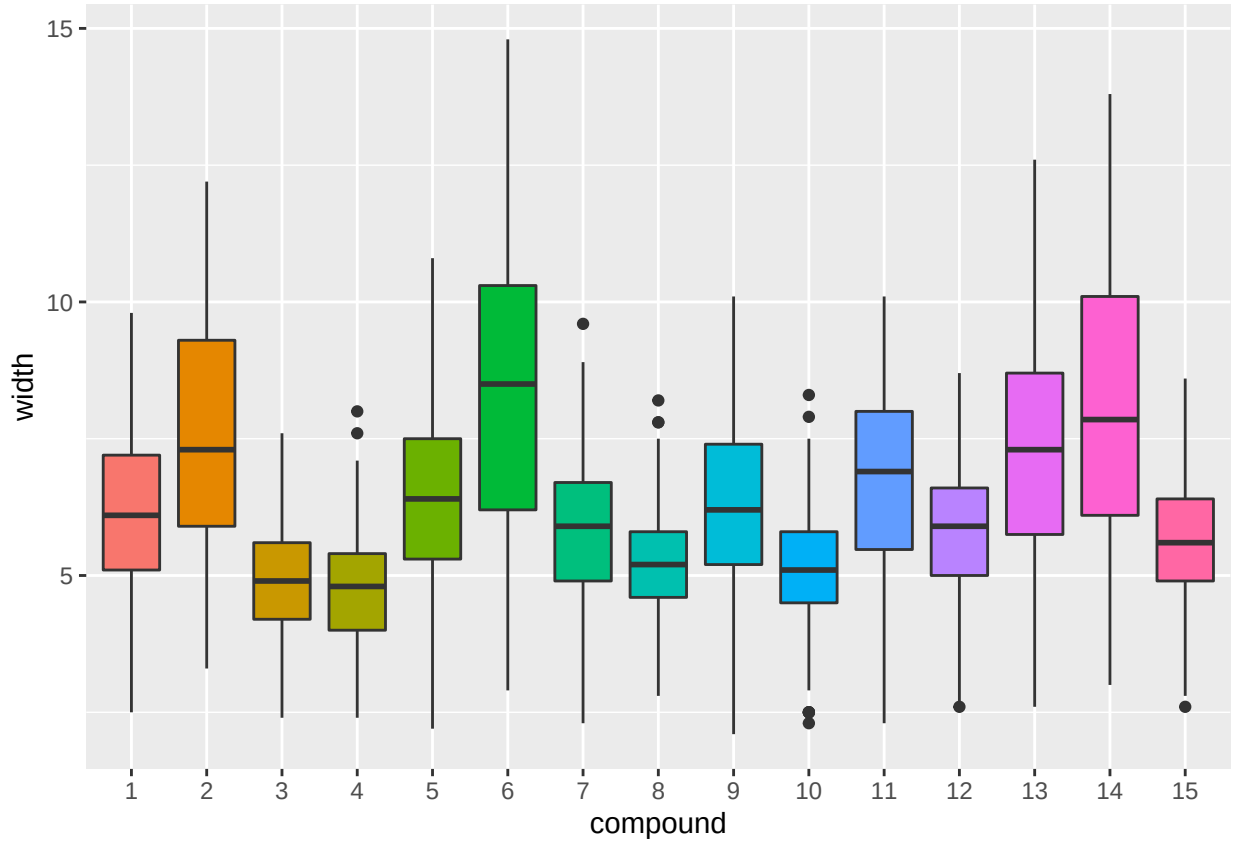

```
ggplot(dataG_combo, aes(x=width, fill=compound))+geom_density(alpha=0.5)+theme(legend.position = "none")
```

```
## Warning: Removed 88 rows containing non-finite values (stat_density).
```




```
ggplot(dataG_combo, aes(x=compound,y=width, fill=compound))+geom_boxplot()+theme(legend.position = "none
```

```
## Warning: Removed 88 rows containing non-finite values (stat_boxplot).
```

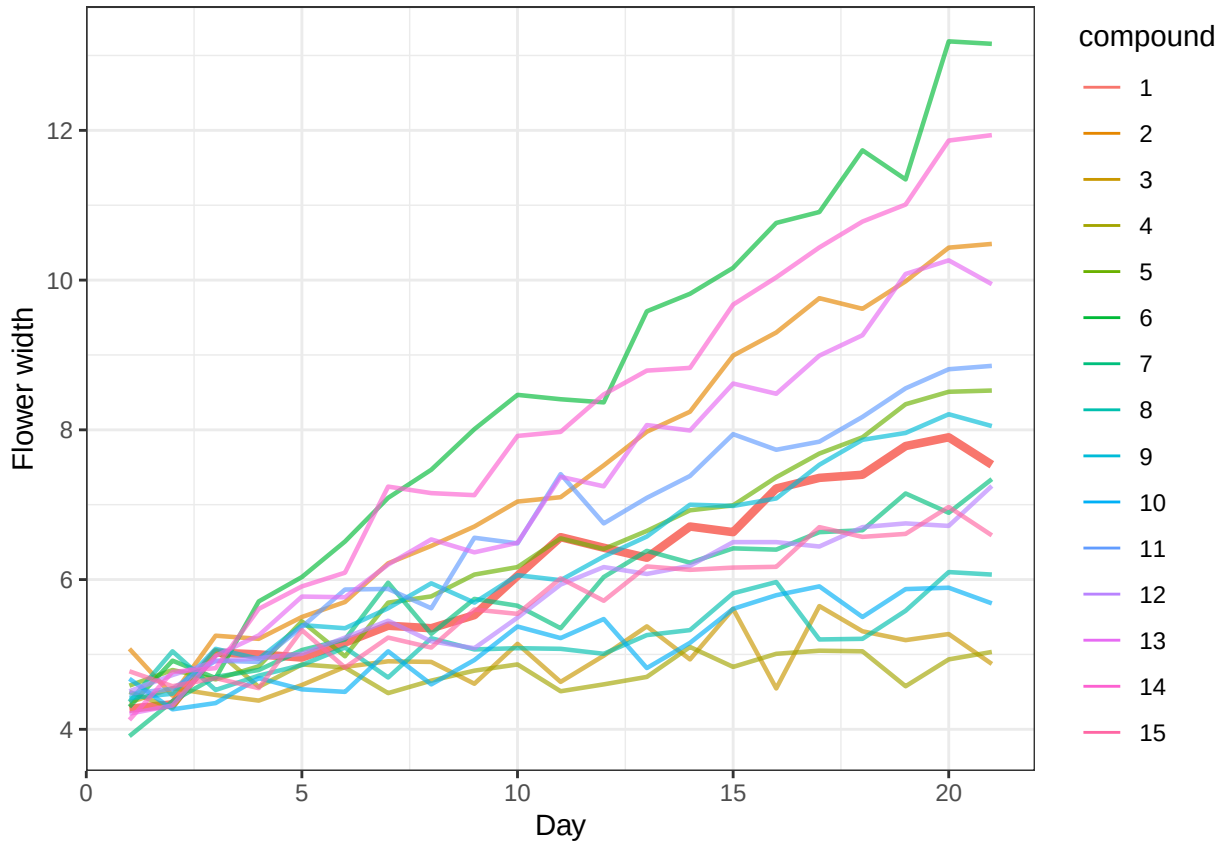


Below I plotted the mean width of the flower by day by compound on a given day.

```
data_cc <- aggregate(width ~ compound + days, data = dataG_combo, FUN = mean) %>%
  mutate(water=ifelse(compound==1,T,F))

ggplot(data = data_cc)+
    geom_line(aes(x = days, y = width, color = compound, size=water, alpha=water)) +
  scale_size_discrete(range=c(0.8,1.5),guide="none")+
  scale_alpha_discrete(range=c(0.65,1), guide="none")+
  theme_bw()+
  ylab("Flower width")+
  xlab("Day")
```

```
## Warning: Using size for a discrete variable is not advised.
```

```
## Warning: Using alpha for a discrete variable is not advised.
```




```
#ggplot(data = data_cc %>% filter(compound %in% c(1, 6, 14)))+
#    geom_line(aes(x = days, y = width, color = compound))
```

The takeaway from this graph is that for each graph, the change in the Width of the flower is not the same for each of the Compounds. Does this mean we have an interaction between Compound and Days?

```
data_ccc<- aggregate(delta_width ~ compound + days, data = dataG_combo, FUN = mean)

plot <- ggplot(data = data_ccc)+
    geom_line(aes(x = days, y = delta_width, color = compound))

plot
```

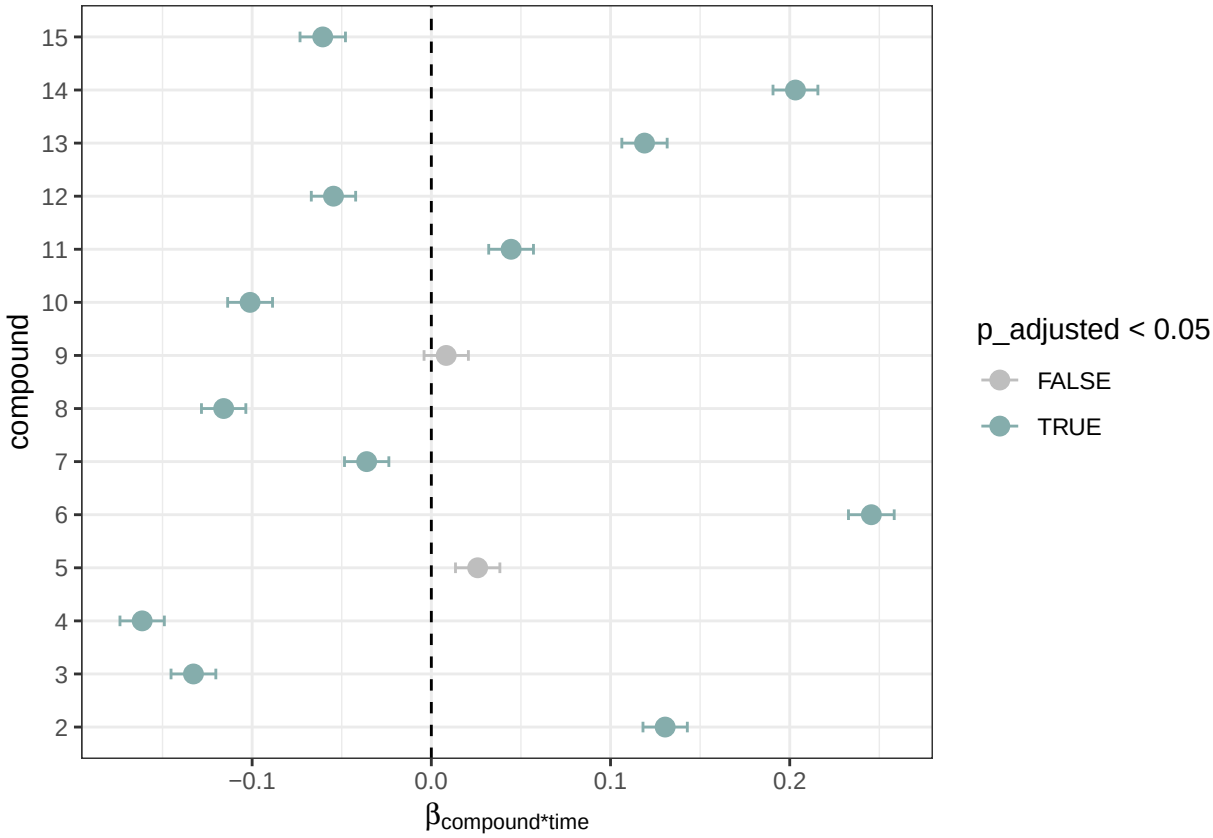
I fit a linear model to the gaussian outcome data where Compound, Type, Garden and Days are included as fixed effects, a compound and days interaction is included and subplot is included as a random effect. Rater is not included because we only have one rater.

```
#g1 <- glm(Width ~ Compound + Type + Garden + Days + Compound*Days + (1 | Subplot), data=dataG_long)
lme_out <- nlme::lme(width ~ compound + type + garden + days + compound*days, data=dataG_long, random =
```

Probably not right, this output is too long.

```
lme_coefficients<-as.data.frame(summary(lme_out)$tTable) %>%
  rownames_to_column("predictor_full") %>%
```

```
  filter(grepl("compound",predictor_full)) %>%
  filter(grepl("days",predictor_full)) %>%
  #dplyr::rename(pval=`p-value`) %>%
  #we want to have p-adjusted (Holm) values for one-sided test H.alt: lambda(compound)>lambda(water)
  #dplyr::mutate(one_sided_pval=ifelse(`t-value`<0, pval/2, (1-pval/2)),
    dplyr::mutate(p_adjusted=p.adjust(`p-value`, method="holm"),
                  significant=ifelse(p_adjusted<0.05, T, F),
                  predictor=gsub(":days","",predictor_full))

ggplot(lme_coefficients %>%
         mutate(compound=factor(gsub("compound|:day","",predictor), levels=2:15)),
       aes(x=compound, y=Value, color=p_adjusted<0.05))+
  geom_hline(yintercept=0, linetype="dashed")+
  geom_errorbar(aes(ymin=Value - `Std.Error`, ymax=Value +`Std.Error`), width=0.2)+geom_point(size=3)+tl
  scale_color_manual(values=c("grey","#85ADAC"))+coord_flip()+
  ylab(expression(beta["compound*time"]))
```

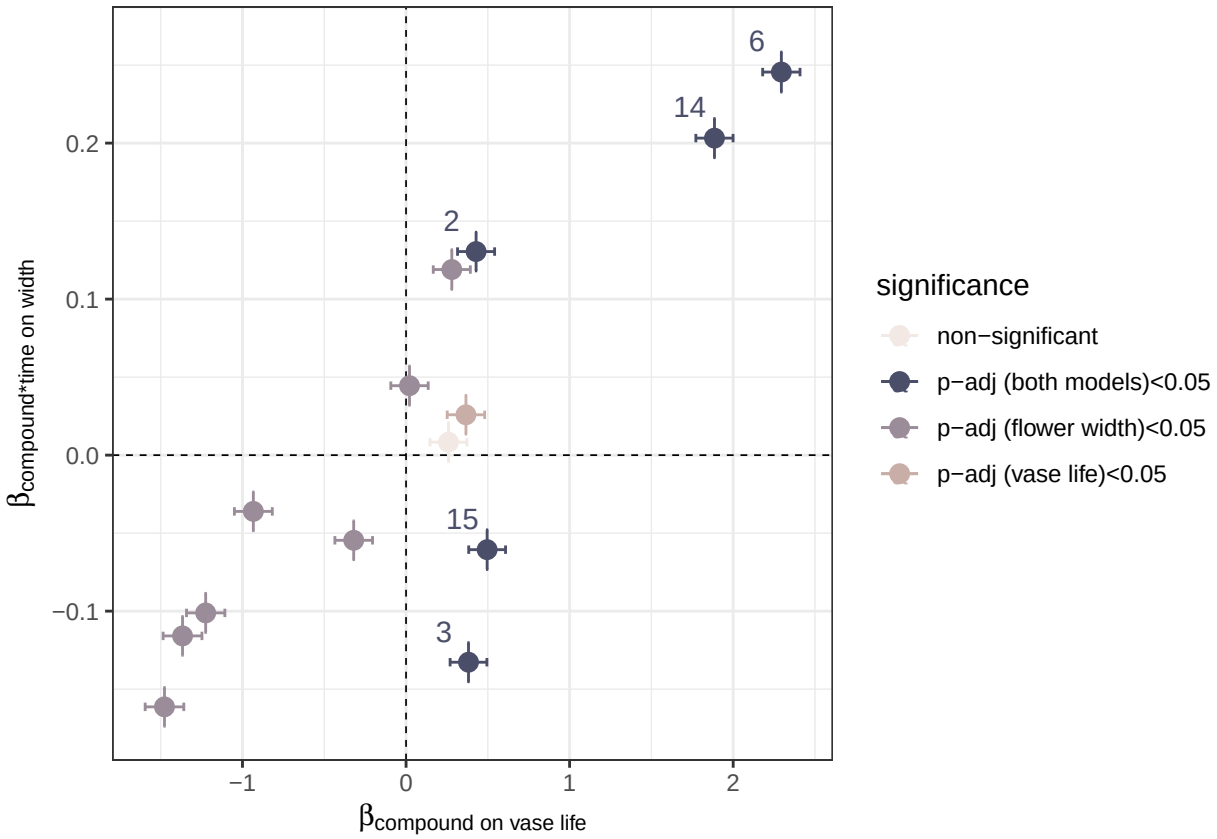



```
# also nice #BEE3DB
```

Now intersect the two model outputs to compare and interpret the results (so far only compared the results of the first and last model).

```
both_predictions<-inner_join(glmer_bn_coefficients, lme_coefficients, by="predictor", suffix=c(".glmer"
  mutate(significance=case_when(p_adjusted.glmer<0.05 & p_adjusted.lme <0.05 ~ "p-adj (both models)<0.05
                                p_adjusted.glmer<0.05 ~ "p-adj (vase life)<0.05",
```

```
                                p_adjusted.lme<0.05 ~ "p-adj (flower width)<0.05",
                                T ~ "non-significant"))
  #mutate(significant_in_either=ifelse(significant_higher | significant, T, F))


ggplot(both_predictions, aes(x=Estimate, y=Value, color=significance))+
  geom_vline(xintercept = 0, linetype="dashed", size=0.3)+
  geom_hline(yintercept = 0, linetype="dashed", size=0.3)+
  geom_point(size=3)+
  geom_errorbarh(aes(xmin=Estimate-`Std. Error`, xmax=Estimate+`Std. Error`))+
  geom_errorbar(aes(ymin=Value-`Std.Error`, ymax=Value+`Std.Error`))+
  xlab("Coefficient (fitted days of vase life)")+
  ylab("Coefficient (fitted slope of flower width over time)")+
  theme_bw()+
  #scale_alpha_discrete(range=c(0.5,1))+
  geom_text(data=subset(both_predictions, significance == "p-adj (both models)<0.05"),
            aes(Estimate,Value,label=gsub("compound","",predictor)), nudge_x = -0.15, nudge_y=0.02)+
  scale_color_manual(values=c("#f2e9e4","#4a4e69","#9a8c98", "#c9ada7"))+
  ylab(expression(beta["compound*time on width"]))+
  xlab(expression(beta["compound on vase life"]))
```



Approximate conclusions

-Count data suggest compounds 6 and 14 as candidates -Longitudinal binary data suggests the same ones, but in addition 3 more, possibly due to higher power that longitudinal modelling provides -Longitudinal

Gaussian data describing width suggests that most compounds besides 2 affect the flower width over time. If the width increase is indeed a good sign, by combining the count data results and these, the compounds 6 and 14 might indeed be good candidates for extending rose vase life relative to water.