Q2

Hun Kang

2024-10-27

Q2.

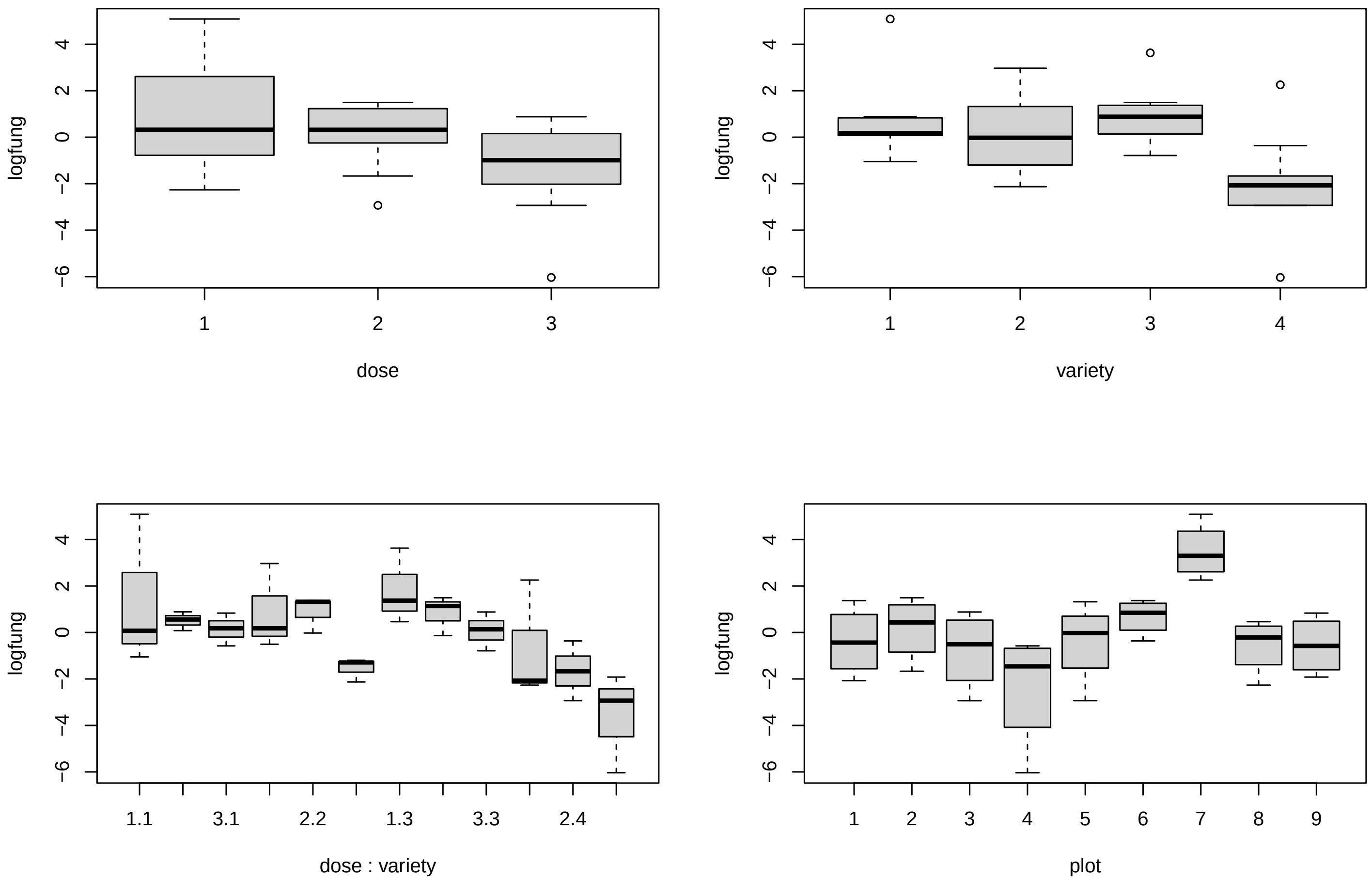
```
library(lme4)
library(tidyverse)
```

```
dat = read.table("fung.dat",header=TRUE)
head(dat)
```

```
##   plot dose variety logfung
## 1    1    1       2   0.178
## 2    1    1       1  -1.049
## 3    1    1       4  -2.074
## 4    1    1       3   1.371
## 5    2    2       4  -1.669
## 6    2    2       1   0.888
```

a.

```
par(mfrow=c(2,2))
boxplot(logfung ~ dose, data = dat)
boxplot(logfung ~ variety, data = dat)
boxplot(logfung ~ dose*variety, data = dat)
boxplot(logfung ~ plot, data = dat)
```



b.

In linear regression, **normality** can be checked by QQ plot, and **(equal) variance** and **independence** assumptions can be typically checked by drawing residuals against some structure in the data, e.g. time or group. If these assumptions are met, the boxplots of residuals per plot should be centered around zero with similar spread. In this data, some plots have higher residuals than others or higher variance (box width). This indicate that there is a pattern in the residuals related to the grouping factor, violating the assumptions.

The p-value of dose is 0.007697, but keep in mind that the model here assumes we have 36 of iid samples. However, as we will see below, if the iid assumption on the error is questionable, than the effective number of iid samples would be smaller. In this aspect we can suspect this p-value is underestimated.

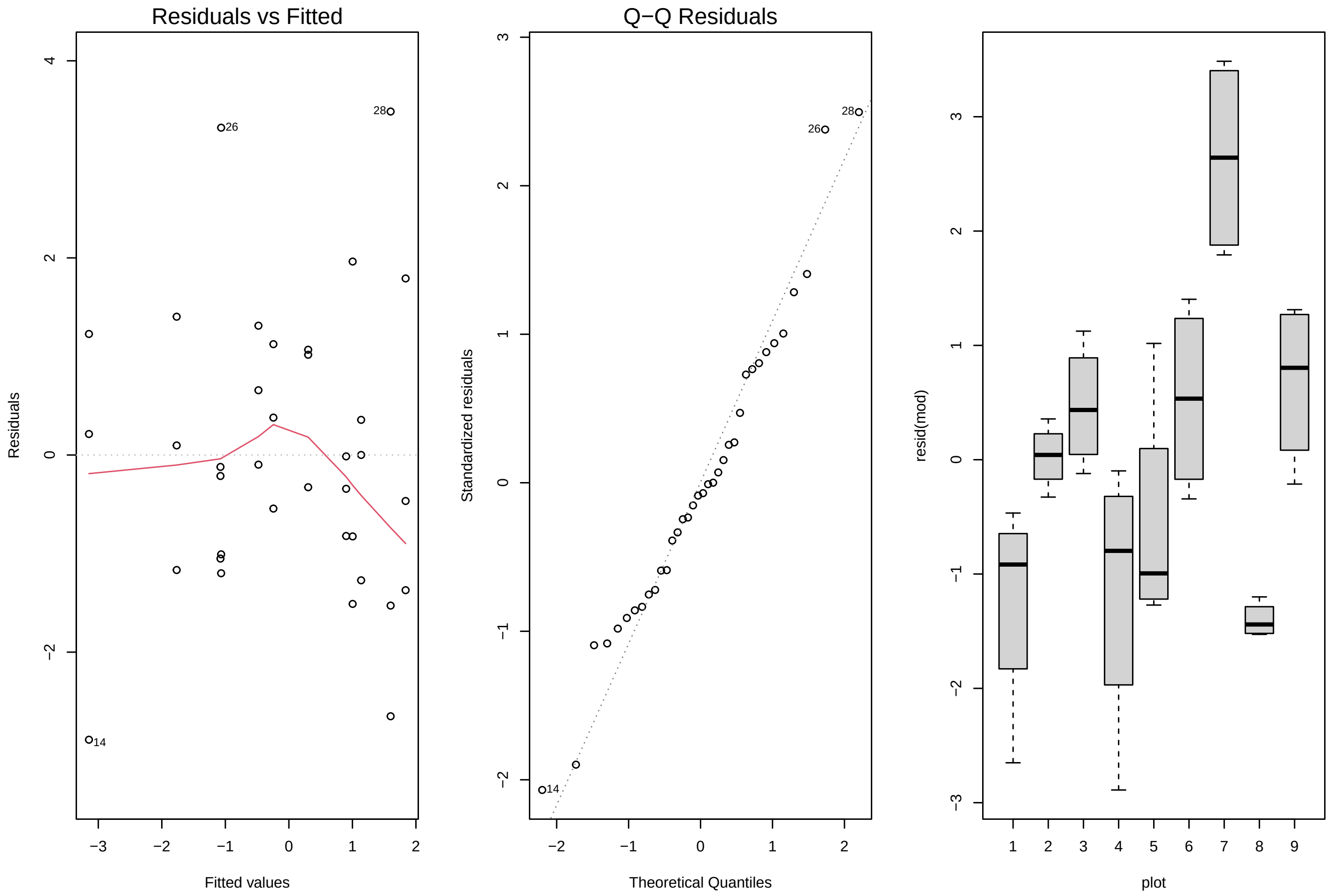
```
## fit the full model
mod = lm(logfung ~ as.factor(dose) + as.factor(variety), data=dat) # no plot effect
summary(mod)
```

```
##
## Call:
## lm(formula = logfung ~ as.factor(dose) + as.factor(variety),
##     data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8888 -1.0186 -0.1094  1.0299  3.4852
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.6018     0.6245   2.565 0.015569 *
## as.factor(dose)2     -0.7001     0.6245  -1.121 0.271200
## as.factor(dose)3     -2.0810     0.6245  -3.332 0.002299 **
## as.factor(variety)2  -0.5979     0.7212  -0.829 0.413615
## as.factor(variety)3   0.2357     0.7212   0.327 0.746097
## as.factor(variety)4  -2.6680     0.7212  -3.700 0.000866 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.53 on 30 degrees of freedom
## Multiple R-squared:  0.5132, Adjusted R-squared:  0.4321
## F-statistic: 6.327 on 5 and 30 DF,  p-value: 0.0004032
```

```r
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: logfung
##                     Df Sum Sq Mean Sq F value   Pr(>F)
## as.factor(dose)      2 26.910 13.4552  5.7493 0.007697 **
## as.factor(variety)   3 47.121 15.7070  6.7115 0.001339 **
## Residuals           30 70.209  2.3403
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
par(mfrow=c(1,3))
plot(mod, which=c(1,2))
boxplot(resid(mod) ~ plot, data = dat)
```



c.

We take mean of each plot, summarizing 36 obs across 9 plots into 9 obs, discarding variety. Therefore the residual variance is the plot to plot variation. Since we are using less data than in the sample, the test will have less power (the probability to reject the null hypothesis if the effect size actually exists), which explains one of why the p-value here is larger than b.

```
datc = dat %>% group_by(plot) %>% summarize(logfung = mean(logfung), dose = first(dose))
datc
```

```
## # A tibble: 9 x 3
##    plot logfung  dose
##   <int>   <dbl> <int>
## 1     1  -0.393     1
## 2     2   0.172     2
## 3     3  -0.768     3
## 4     4  -2.38      3
## 5     5  -0.416     2
## 6     6   0.676     2
## 7     7   3.48      1
## 8     8  -0.558     1
## 9     9  -0.56      3
```

```
mod = lm(logfung ~ as.factor(dose), data = datc)
summary(mod)
```

```
## 
## Call:
## lm(formula = logfung ~ as.factor(dose), data = datc)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.40250 -1.14525  0.02833  0.53233  2.64025 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.8442     0.8517   0.991    0.360
## as.factor(dose)2  -0.7001     1.2045  -0.581    0.582
## as.factor(dose)3  -2.0810     1.2045  -1.728    0.135
## 
## Residual standard error: 1.475 on 6 degrees of freedom
## Multiple R-squared:   0.34,  Adjusted R-squared:   0.12 
## F-statistic: 1.546 on 2 and 6 DF,  p-value: 0.2875
```

```
anova(mod)
```

```
## Analysis of Variance Table
## 
## Response: logfung
##                 Df  Sum Sq Mean Sq F value Pr(>F)
## as.factor(dose)  2  6.7276  3.3638  1.5457 0.2875
## Residuals        6 13.0576  2.1763
```

d

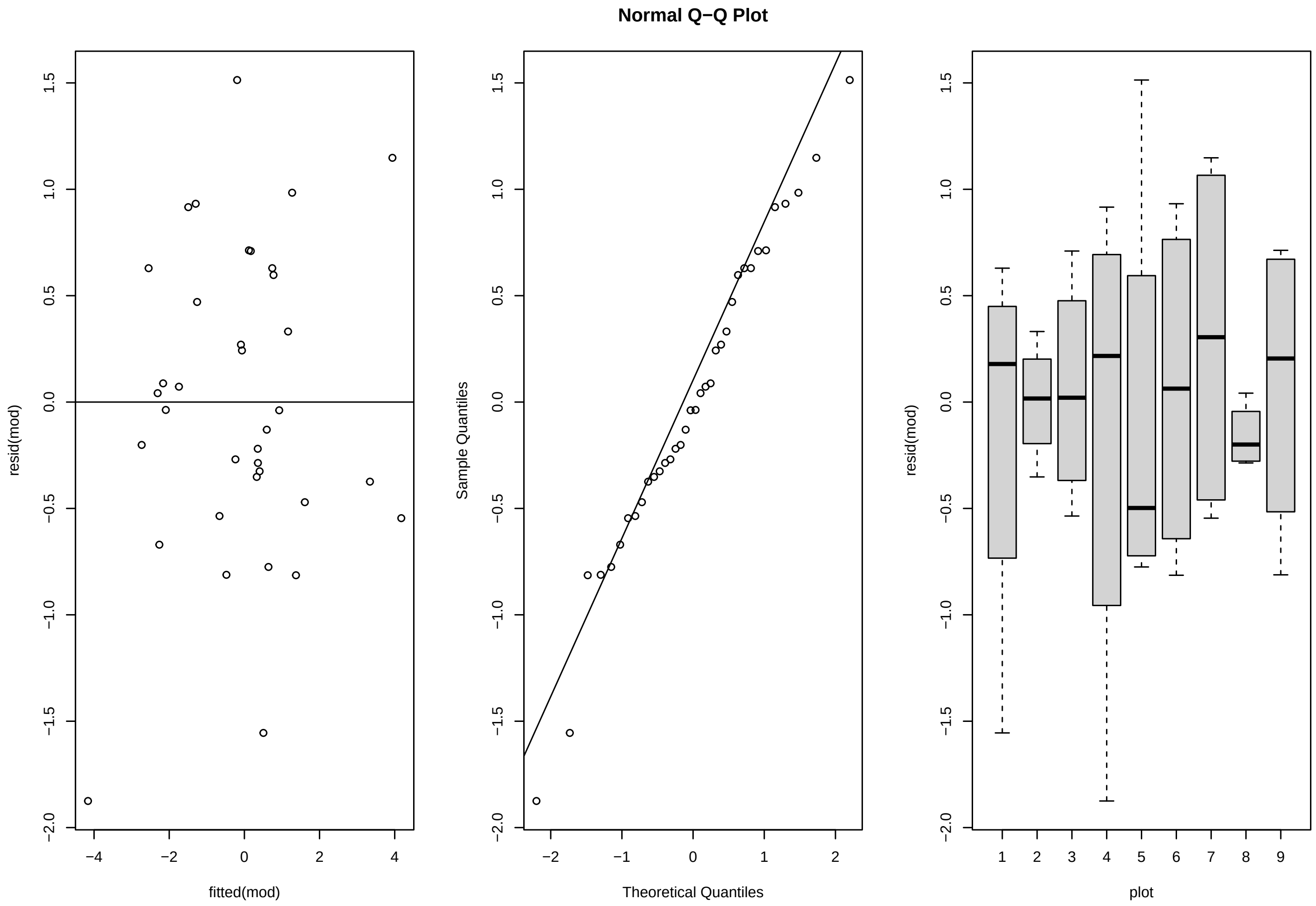
After including the plot random intercept, the reisuals per group appear better than b. The p-value here is based on the asymptotic distribution of the likelihood ratio test statistics, which is Chi-squared. This is the distribution of the test statistics we would see if take all the predictors (dose, variety and plot), X as given, and obtain imaginary samples of the logfung, y, under the null hypothesis that the regression coefficients of dose dummy variables are zero. The p-value here is larger than b without considering plot and smaller than c which only uses 9 summarized data.

```
mod = lmer(logfung ~ as.factor(dose) + as.factor(variety) + (1|plot), data = dat, REML=F)
summary(mod)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: logfung ~ as.factor(dose) + as.factor(variety) + (1 | plot)
##    Data: dat
## 
```

```
##      AIC      BIC   logLik deviance df.resid
##    123.0    135.7    -53.5    107.0       28
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.29768 -0.48812 -0.04649  0.74135  1.85472
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  plot     (Intercept) 1.2844   1.133
##  Residual             0.6659   0.816
## Number of obs: 36, groups:  plot, 9
##
## Fixed effects:
##                      Estimate Std. Error t value
## (Intercept)            1.6018     0.7342   2.182
## as.factor(dose)2      -0.7001     0.9835  -0.712
## as.factor(dose)3      -2.0810     0.9835  -2.116
## as.factor(variety)2  -0.5979     0.3847  -1.554
## as.factor(variety)3   0.2357     0.3847   0.613
## as.factor(variety)4  -2.6680     0.3847  -6.936
##
## Correlation of Fixed Effects:
##             (Intr) as.fctr(d)2 as.fctr(d)3 as.fctr(v)2 as.fctr(v)3
## as.fctr(d)2 -0.670
## as.fctr(d)3 -0.670  0.500
## as.fctr(v)2 -0.262  0.000       0.000
## as.fctr(v)3 -0.262  0.000       0.000       0.500
## as.fctr(v)4 -0.262  0.000       0.000       0.500       0.500
```

```
par(mfrow=c(1,3))
plot(fitted(mod), resid(mod)); abline(h=0)
qqnorm(resid(mod)); qqline(resid(mod))
boxplot(resid(mod) ~ plot, data = dat)
```



```
mod_nodose = lmer(logfung ~ as.factor(variety) + (1|plot), data = dat, REML=F)
(res = anova(mod, mod_nodose))
```

```
## Data: dat
## Models:
## mod_nodose: logfung ~ as.factor(variety) + (1 | plot)
## mod: logfung ~ as.factor(dose) + as.factor(variety) + (1 | plot)
##            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## mod_nodose    6 122.75 132.25 -55.375   110.75
## mod           8 123.01 135.68 -53.505   107.01 3.7401  2     0.1541
```

```
(Chisq_obs = res$Chisq[2])
```

```
## [1] 3.740088
```

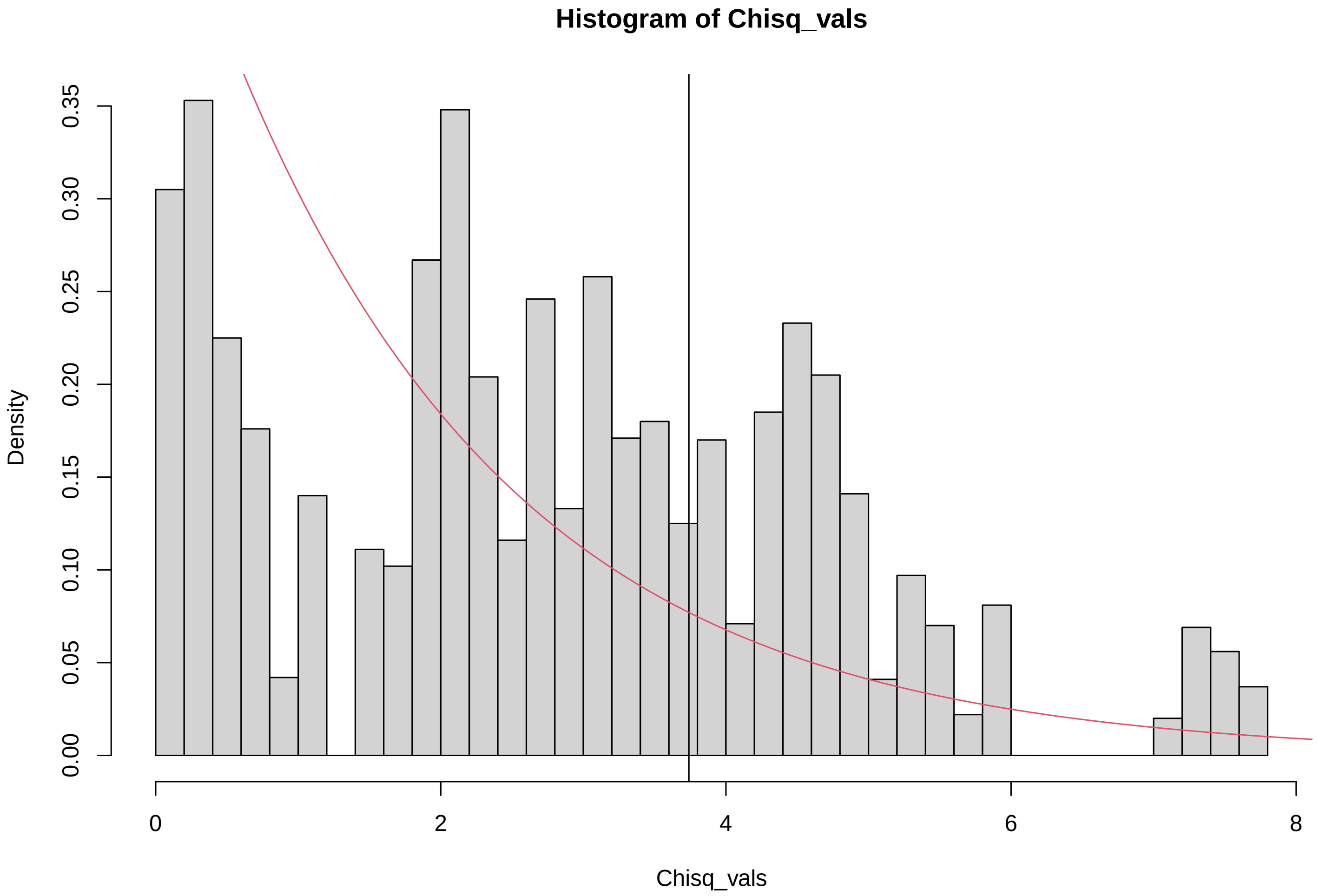
e.

Unlike in d where we fix X and sample y under the null hypothesis, here in the randomized test, we fix y and sample X under the null hypothesis. If the dose indeed has no effect at all, then even if we change the dose assignment, it will not change the response y. By randomly assigning dose to each plot and computing the LRT test statistic, we can obtain another null distribution of the test statistics. The difference is that in d, the randomness of the null distribution comes from y: another set of logfung we would observe under the same dose allocation if the dose had no effect, whereas in e, the randomness comes from randomly assigning dose allocation, X. These two are different approaches in obtaining a null distribution of the test statistics, and has no reason to be equal at all. The p-value here is about 0.3.

```
nsim = 5000
datsim = dat
Chisq_vals = numeric(nsim)
for(i in 1:nsim){
  if(i %% 500 == 0) print(i)
  datsim$dose = rep(sample(rep(1:3, 3)), each = 4)
  mod = lmer(logfung ~ as.factor(dose) + as.factor(variety) + (1|plot), data = datsim, REML=F)
  mod_nodose = lmer(logfung ~ as.factor(variety) + (1|plot), data = datsim, REML=F)
  res = anova(mod, mod_nodose)
  Chisq_vals[i] = res$Chisq[2]
}
```

```
## [1] 500
## [1] 1000
## [1] 1500
## [1] 2000
## [1] 2500
## [1] 3000
## [1] 3500
## [1] 4000
## [1] 4500
## [1] 5000
```

```
hist(Chisq_vals, breaks=30, freq=F)
x = seq(0, 9, len=200)
y = dchisq(x, 2)
abline(v = Chisq_obs)
lines(x, y, col=2)
```



```
mean(Chisq_vals > Chisq_obs)
```

```
## [1] 0.303
```