Bootstrap ANOVA

ANOVA using the median absolute deviations

Solon Karapanagiotis

3 August 2016

The `plyr` and `car` packages need to be installed.

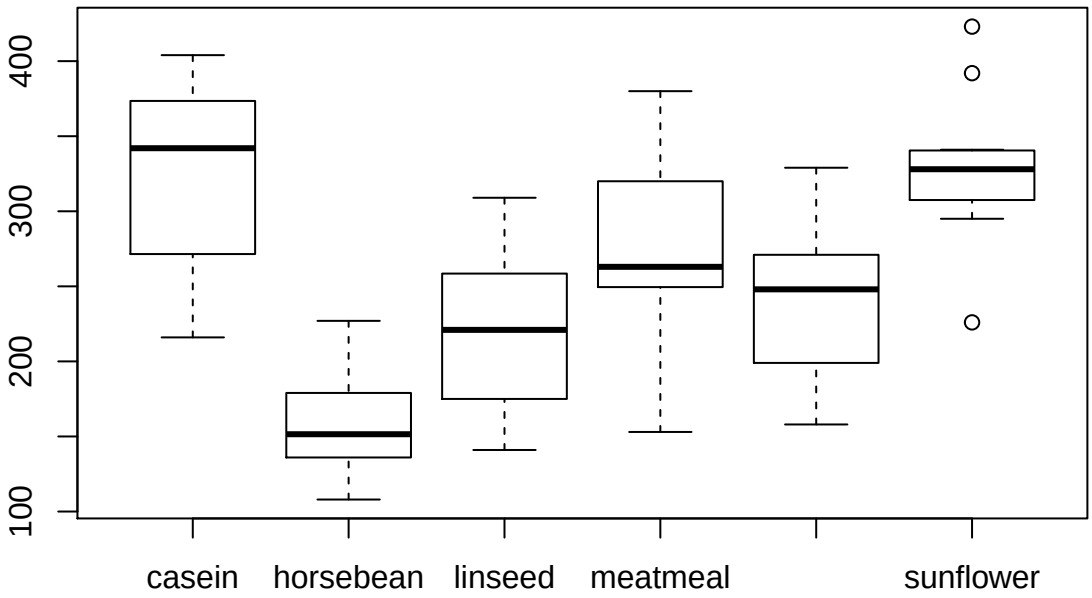
The data I use is the Chicks dataset. The dataset contains information over an experiment conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Their weights (in grams) after six weeks are given along with the feed type. The data-frame contains 71 observations.

```
data(chickwts)
str(chickwts)
```

```
## 'data.frame':    71 obs. of  2 variables:
##  $ weight: num  179 160 136 227 217 168 108 124 143 140 ...
##  $ feed  : Factor w/ 6 levels "casein","horsebean",..: 2 2 2 2 2 2 2 2 2 2 ...
```

The figure shows the distribution of weights by feed supplement. We see that the distribution of the weights is fairly symmetric for each diet group and the median values seem to fluctuate substantially. The same is valid looking at the means (table below). Chicken that were fed with horsebean exhibited the smaller weight on average. The largest value is 328 grams, for those on the sunflower diet.

```
boxplot(weight~feed, data=chickwts)
```



```r
library(plyr)
ddply(chickwts, c("feed"), summarise,
      N = length(weight), mean = round(mean(weight),2),
      sd = round(sd(weight),2))
```

```
##        feed  N   mean    sd
## 1    casein 12 323.58 64.43
## 2 horsebean 10 160.20 38.63
## 3   linseed 12 218.75 52.24
## 4  meatmeal 11 276.91 64.90
## 5   soybean 14 246.43 54.13
## 6 sunflower 12 328.92 48.84
```

The question of interest is if there a difference between the chick weights across the diet groups. In hypothesis framework this can be formulated as

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_6$$

$$H_a : \text{at least one pair of } \mu_1, \ldots, \mu_6 \text{ are different}$$

using μ_1 to denote the mean weight of chickens fed on casein, μ_2 to denote the mean weight of chickens fed on horsebean etc. To test H_0 the one-way ANOVA model can be formulated

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

where Y_{ij} is the weight of chick i in diet group j, μ_i are the parameters and ϵ_{ij} are the error terms. The ANOVA table is given

```r
oneway.test(weight ~ feed, data=chickwts, var.equal = T)
```

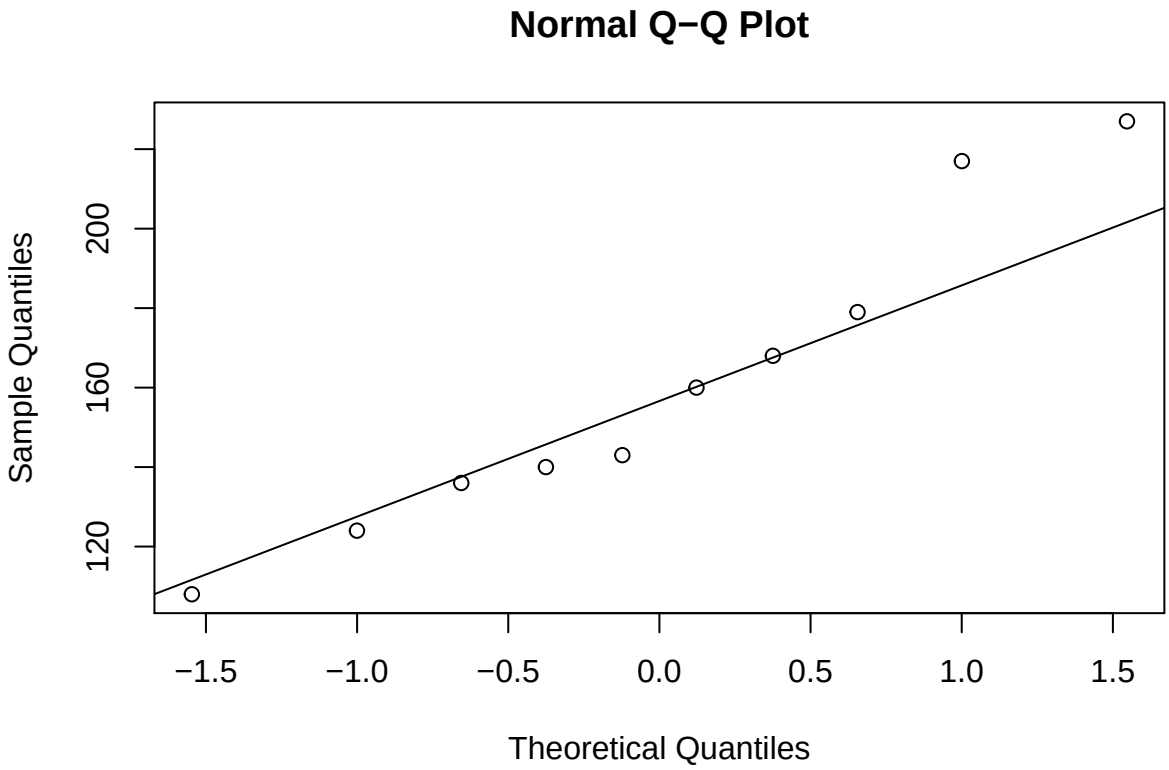
```
##
##  One-way analysis of means
##
## data:  weight and feed
## F = 15.365, num df = 5, denom df = 65, p-value = 5.936e-10
```

- with the `var.equal = T` statement I treat the variances in the samples as equal. Motivation follows in the *Comments* below.

When H_0 holds the F statistic is distributed as $F(5, 65) = 2.35$. The p-value is the probability $P\big(F(5, 65) > F_{stat} = 15.37\big) = 5^{-10}$. Based on this value we conclude that the data are not consistent with the null hypothesis. Hence, we can say that there are differences between the mean weights across the feed supplements (at 5% significance level).

Comments: The ANOVA model formulated above assumes that each probability distribution (for each diet group) is normal and they have the same variance (homoskedasticity). For the normality assumption we created the quantile-quantile plots for each feed type. For example, for the horsebean

```r
# normality: Each probability distribution is normal
qqnorm(chickwts$weight[chickwts$feed == "horsebean"])
qqline(chickwts$weight[chickwts$feed == "horsebean"])
```



Similar plots can be created for the other feed categories. Overall, not substantial deviations from this assumption were noticed. I tested the homoskedasticity assumption using the Levene's test (see code). We have no evidence (p=0.58), at the 5% level, that the variances are not homogeneous. Hence, the choice of `var.equal = T` previously.

```
library(car)
leveneTest(chickwts$weight ~ chickwts$feed)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  5  0.7493 0.5896
##       65
```

In the ANOVA calculation above the F statistic was based on the ratio of "between group" (SSTR=231129) and "within group" (SSE=195556) sum of squares (adjusted to the degree of freedoms).

```
summary(aov(weight~feed, data=chickwts))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## feed         5 231129   46226   15.37 5.94e-10 ***
## Residuals   65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The `aov()` is just another fuction to perform ANOVA. The results are the same, the output changes. For more background into ANOVA my favourite book is: Applied linear statistical models by Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005).

Now, I formulate three other statistics, the sum of absolute deviances. I define the total sum of absolute deviances (ADTO):

$$ADTO = \sum_i \sum_j |Y_{ij} - M_{..}|$$

The *treatment sum of absolute deviances (ADTR)*:

$$ADTR = \sum_i n_i |M_{i.} - M_{..}|$$

The *error sum of absolute deviances (ADSE)*:

$$ADSE = \sum_i \sum_j |Y_{ij} - M_{i.}|$$

where $M_{i.}$ is the median over the of the j'th diet group and $M_{..}$ the overall median. The difference with the classical ANOVA is that the mean is substitued by the median and the sum of squared deviances by the sum of absolute deviances.

```
##ADTO: total sum of absolute deviances
ADTO <- sum(abs(chickwts$weight-median(chickwts$weight)))

##ADSE error sum of absolute deviances
ADSE <- sum(abs(chickwts$weight[chickwts$feed == "horsebean"]- median(chickwts$weight[chickwts$feed ==
  sum(abs(chickwts$weight[chickwts$feed == "linseed"]- median(chickwts$weight[chickwts$feed == "linseed
  sum(abs(chickwts$weight[chickwts$feed == "soybean"]- median(chickwts$weight[chickwts$feed == "soybean
  sum(abs(chickwts$weight[chickwts$feed == "sunflower"]- median(chickwts$weight[chickwts$feed == "sunfl
  sum(abs(chickwts$weight[chickwts$feed == "meatmeal"]- median(chickwts$weight[chickwts$feed == "meatmea
  sum(abs(chickwts$weight[chickwts$feed == "casein"]- median(chickwts$weight[chickwts$feed == "casein"])

#ADTR: treatment absolute deviances
ADTR <- ADTO-ADSE

MADTR <- ADTR/5 #r-1 degress of freedom, r for the groups
MADE <- ADSE/(length(chickwts$weight)-6) #n-r degrees of freedom

F_tilde_star = MADTR/MADE
```

The table summarizes the above quantities

	DF	AD_	MAD_	$\tilde{F}$
Feed	5	1654	330.6	7.29
Residuals	65	2944	45.29	

The $\tilde{F}$ statistic is the ratio $\frac{MADTR}{MADSE} = \frac{330.6}{45.29} = 7.29$.

Given the construction of the above statistics as deviations from the median a natural hypothesis to test is:

$$H_0 : m_1 = m_2 = \ldots = m_6 \tag{1}$$

$$H_a : \text{at least one pair of } m_1, \ldots, m_6 \text{ are different}$$

where m_1 to denote the median weight of chickens fed on casein, m_2 to denote the median weight of chickens fed on horsebean etc.

Since we do not know the theoretical distribution of $\tilde{F}$ we will use parametric bootstrap to simulate the theoretical distribution $\tilde{F}^*$ under the null hypothesis. To keep in line with the assumptions of the classical ANOVA since they were not rejected (see *Comments* above) I assume that each Y_{ij} comes from a normal distribution with median=258 and standard deviation=54.6: $Y_{ij} \sim N(258, 54.6)$. The median corresponds to the median of the whole dataset. Since we assume normal distribution mean=median. The standard deviation is the pooled SD from the dataset. I choose to use the pooled value based on the Levene's test, where heteroskedasticity was rejected.

```
median(chickwts$weight)
```

```
## [1] 258
```

```
# since can reject homogeneity we assume pooled sd
sqrt(mean(with(chickwts,tapply(weight,feed,var))))
```

```
## [1] 54.6188
```

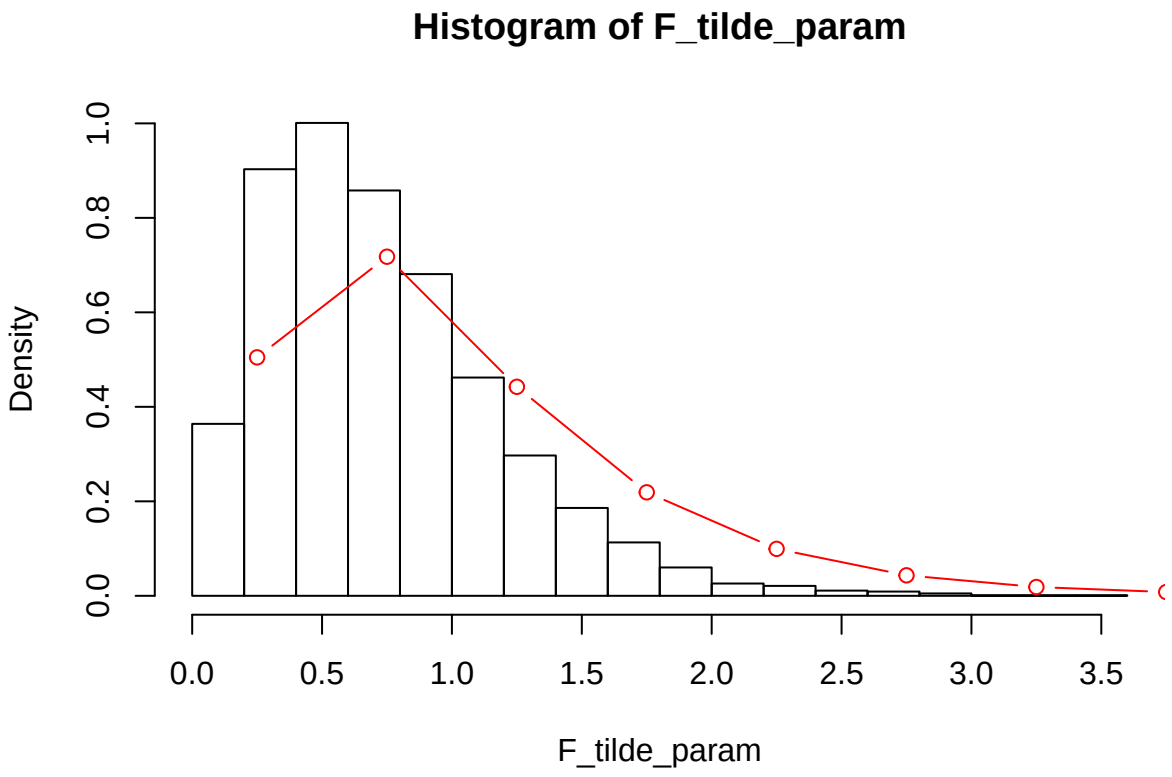
The bootstrap algorithm proceeds as follows:
1. n_t values for Y_{ij} are sampled from $N(258, 54.6)$ (n_t is the number of chicks in each feed category),
2. $ADTR$, $ADSE$, $MADTR$, $MADE$ and $\tilde{F}^*$ are computed
3. steps 1 and 2 are repeated R=5000 times

```
## parametric bootstrap
medianstar = median(chickwts$weight)  #the median (=mean) value of the nomral distribution
sdstar = sqrt(mean(with(chickwts,tapply(weight,feed,var))))   #the pooled sd value of the nomral distri
simfeed = chickwts$feed   # the feed categories
R = 5000  # number of bootstrap replications
ADTO <- numeric(R) # empty vectors of the quantities I need. They will hold the simulated values later.
ADSE <- numeric(R)
ADTR <- numeric(R)
MADE <- numeric(R)
MADTR <- numeric(R)
F_tilde_param = numeric(R)
set.seed(23455)

#the groups A-F are not needed here but I left for convenience (I used them in the non-paramtric later)
for (i in 1:R) {
  groupA = rnorm(12, mean=medianstar, sd=sdstar) # 6 groups, one for each feed supplement
  groupB = rnorm(10, mean=medianstar, sd=sdstar)
  groupC = rnorm(12, mean=medianstar, sd=sdstar)
  groupD = rnorm(11, mean=medianstar, sd=sdstar)
  groupE = rnorm(14, mean=medianstar, sd=sdstar)
  groupF = rnorm(12, mean=medianstar, sd=sdstar)
  sim_weight = c(groupA,groupB,groupC,groupD,groupE,groupF)
  simdata = data.frame(sim_weight,simfeed)

  ADTO[i] <- sum(abs(simdata$sim_weight-median(simdata$sim_weight))) #the index i means that I calculat
```

```
  ## ADSE error sum of absolute deviances
  ADSE[i] <- sum(abs(simdata$sim_weight[simdata$simfeed == "horsebean"]- median(simdata$sim_weight[simda
    sum(abs(simdata$sim_weight[simdata$simfeed == "linseed"]- median(simdata$sim_weight[simdata$simfeed
    sum(abs(simdata$sim_weight[simdata$simfeed == "soybean"]- median(simdata$sim_weight[simdata$simfeed
    sum(abs(simdata$sim_weight[simdata$simfeed == "sunflower"]- median(simdata$sim_weight[simdata$simfe
    sum(abs(simdata$sim_weight[simdata$simfeed == "meatmeal"]- median(simdata$sim_weight[simdata$simfeed
    sum(abs(simdata$sim_weight[simdata$simfeed == "casein"]- median(simdata$sim_weight[simdata$simfeed =

  #ADTR: treatment absolute deviances
  ADTR[i] <- ADTO[i]-ADSE[i]

  MADTR[i] <- ADTR[i]/5 #r-1 degress of freedom, r for the groups
  MADE[i] <- ADSE[i]/(length(chickwts$weight)-6) #n-r degrees of freedom

  F_tilde_param[i] = MADTR[i]/MADE[i] #final step calculate the F_tilde for each replication

}
```

The $\tilde{F}^*$ is presented in the figure. The red line corresponds to $F(5, 65)$. Notice that the theoretical distribution used in the classical ANOVA does not correspond to the simulated one. $\tilde{F}^*$ is heavier in the tails. This indicates that if we had used it as the theoretical distribution of $\tilde{F}^*$ we would have obtained a p-value equal to 1^{-5}. This is about 10 times smaller than the Monte-Carlo p-value.

```
hist(F_tilde_param, prob=T)
x=seq(.25,5.25,.5)
points(x,y=df(x,5,66),type="b",col="red")
```



The Monte-Carlo p-value is calculated as

$$p = \frac{1 + \#(\tilde{F}^* > \tilde{F})}{R + 1}$$

```
(p_value <- (1+sum(F_tilde_param > F_tilde_star))/(R+1))
```

```
## [1] 0.00019996
```

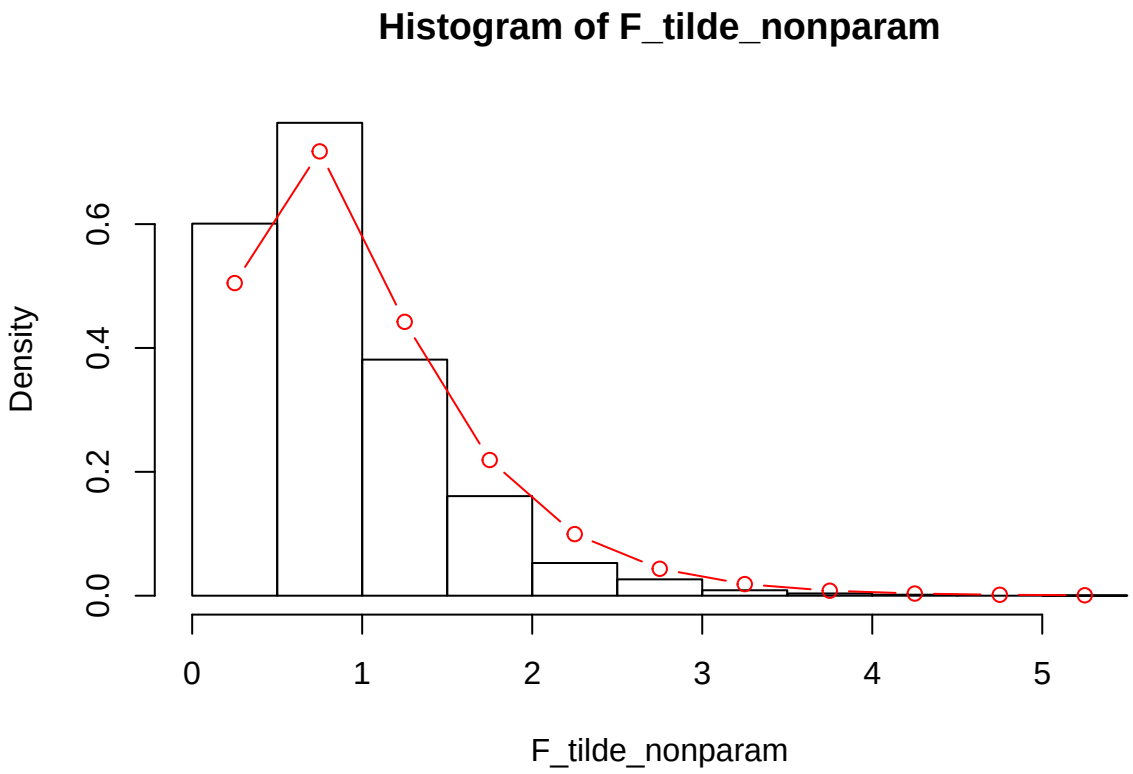
The small p-value allows us to reject the null hypothesis and conclude the median weights are different for each diet group of chickens.

Another alternative is not to assume a parametric model for Y_{ij}. In that case the bootstrap algorithm proceeds as follows:
1. first I center all the groups on the same median (zero), in order to simulate under H_0, and left the variance and shape of the individual group distributions undisturbed
2. n_t values for Y_{ij} are sampled from the empirical distributions
3. $ADTR$, $ADSE$, $MADTR$, $MADE$ and $\tilde{F}^*$ are computed
4. steps 2 and 3 are repeated 5000 times

```
## Non-parametric bootstrap
medianstar = with(chickwts, tapply(weight,feed,median))
grpA = chickwts$weight[chickwts$feed == "horsebean"] - medianstar["horsebean"]
grpB = chickwts$weight[chickwts$feed == "linseed"] - medianstar["linseed"]
grpC = chickwts$weight[chickwts$feed == "soybean"] - medianstar["soybean"]
grpD = chickwts$weight[chickwts$feed == "sunflower"] - medianstar["sunflower"]
grpE = chickwts$weight[chickwts$feed == "meatmeal"] - medianstar["meatmeal"]
grpF = chickwts$weight[chickwts$feed == "casein"] - medianstar["casein"]
simfeed = chickwts$feed
R = 5000
F_tilde_nonparam <- numeric(R)
ADTO <- numeric(R)
ADSE <- numeric(R)
ADTR <- numeric(R)
MADE <- numeric(R)
MADTR <- numeric(R)

set.seed(23456)
for (i in 1:R) {
  groupA = sample(grpA, size=10, replace=T)
  groupB = sample(grpB, size=12, replace=T)
  groupC = sample(grpC, size=14, replace=T)
  groupD = sample(grpD, size=12, replace=T)
  groupE = sample(grpE, size=11, replace=T)
  groupF = sample(grpF, size=12, replace=T)
  sim_weight = c(groupA,groupB,groupC,groupD,groupE,groupF)
  simdata = data.frame(sim_weight,simfeed)

  ADTO[i] <- sum(abs(simdata$sim_weight-median(simdata$sim_weight)))

  ##ADSE error sum of absolute deviances
  ADSE[i] <- sum(abs(simdata$sim_weight[simdata$simfeed == "horsebean"]- median(simdata$sim_weight[simda
    sum(abs(simdata$sim_weight[simdata$simfeed == "linseed"]- median(simdata$sim_weight[simdata$simfeed
    sum(abs(simdata$sim_weight[simdata$simfeed == "soybean"]- median(simdata$sim_weight[simdata$simfeed
```

```
    sum(abs(simdata$sim_weight[simdata$simfeed == "sunflower"]- median(simdata$sim_weight[simdata$simfe
    sum(abs(simdata$sim_weight[simdata$simfeed == "meatmeal"]- median(simdata$sim_weight[simdata$simfeed
    sum(abs(simdata$sim_weight[simdata$simfeed == "casein"]- median(simdata$sim_weight[simdata$simfeed =

  #ADTR: treatment absolute deviances
  ADTR[i] <- ADTO[i]-ADSE[i]

  MADTR[i] <- ADTR[i]/5 #r-1 degress of freedom, r for the groups
  MADE[i] <- ADSE[i]/(length(chickwts$weight)-6) #n-r degrees of freedom

  F_tilde_nonparam[i] = MADTR[i]/MADE[i]

}
```

```
hist(F_tilde_nonparam, prob=T)
x=seq(.25,5.25,.5)
points(x,y=df(x,5,65),type="b",col="red")
```




```
(pvalue <- (1+sum(F_tilde_nonparam > F_tilde_star))/(R+1))
```

```
## [1] 0.00019996
```

The p-value=0.0002, the same as before. A possible explanation is because both under the parametric and non-parametric model the $\tilde{F}^*$ is almost flat at the right of $\tilde{F}$. A similar histogram as before is presented. We now see that the simulated distribution is again heavier in the tail than the $F(5, 65)$ but in a lesser extend than under the parametric bootstrap.

Summarising,

- $\tilde{F^*}_{parametric}$ is based on equal medians (the null hypothesis) under the nomrality and homogeneity assumption
  - $\tilde{F^*}_{non-parametric}$ is based on equal medians (the null hypothesis), but normality and homogeneity are no longer assumed
  - $F(5, 35)$ is the theoretical distribution under the null hypothesis using the classical ANOVA.

The three distributions are pictured in the figure below. First of all, we notice that the $\tilde{F^*}_{parametric}$ and $\tilde{F^*}_{non-parametric}$ are very similar to each other. This can be considered as evidence in favour of the parametric assumptions we adopted earlier. Secondly, we could safely say that all of three would give relatively the same result for a statistic > 3. But in earlier instances probably the p-values would be considerably divergent. This divergence would possibly be more evident between the parametric $\tilde{F^*}$ and $F(5, 65)$. In any case the difference in the p-value will depend highly on the statistic since the relative discrepancy of the three distributions is different for different values of the x-axis.

To conclude, it would be inappropriate to use the theoretical results of the classical ANOVA to test the hypothesis about medians. Even though the results would be qualitatively the same in our example this may not be the case in other scenarios.

```
y <- seq(0,5.25,.05)
plot(density(F_tilde_param), col="dark green", main="Desnity plot of distribution of F",lwd=2)
lines(density(F_tilde_nonparam), col="blue", lwd=2)
points(y,y=df(y,5,66), type="l", col="red", lwd=2)
legend(2.5,0.8, c("param", "non-param", "(5,65)"), lty=1, col=c("dark green", "blue", "red"), cex=1.5,lw
```