Stats 330 A3

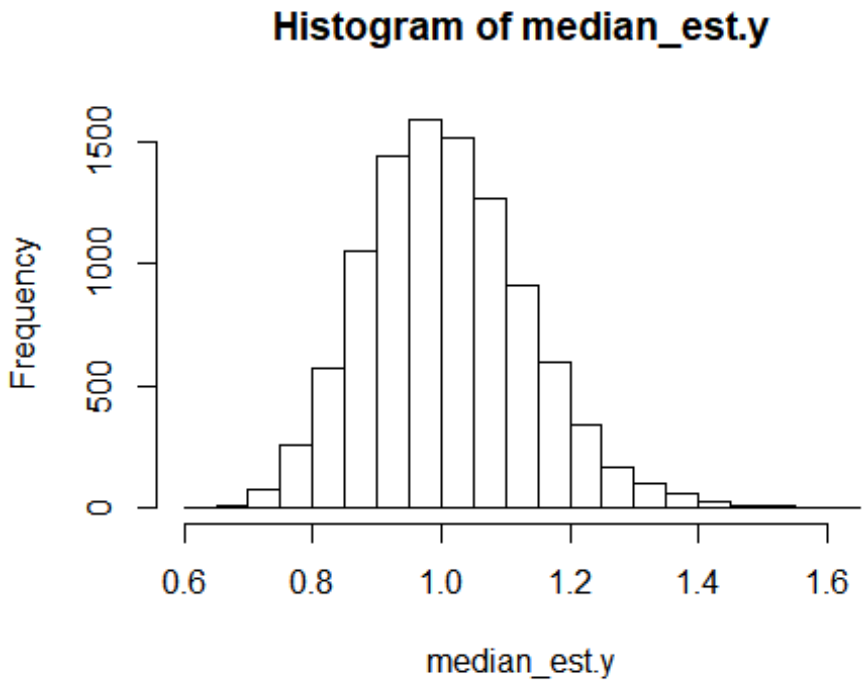
Richard Choi 915726645

Due Date: 12pm Friday 11 October

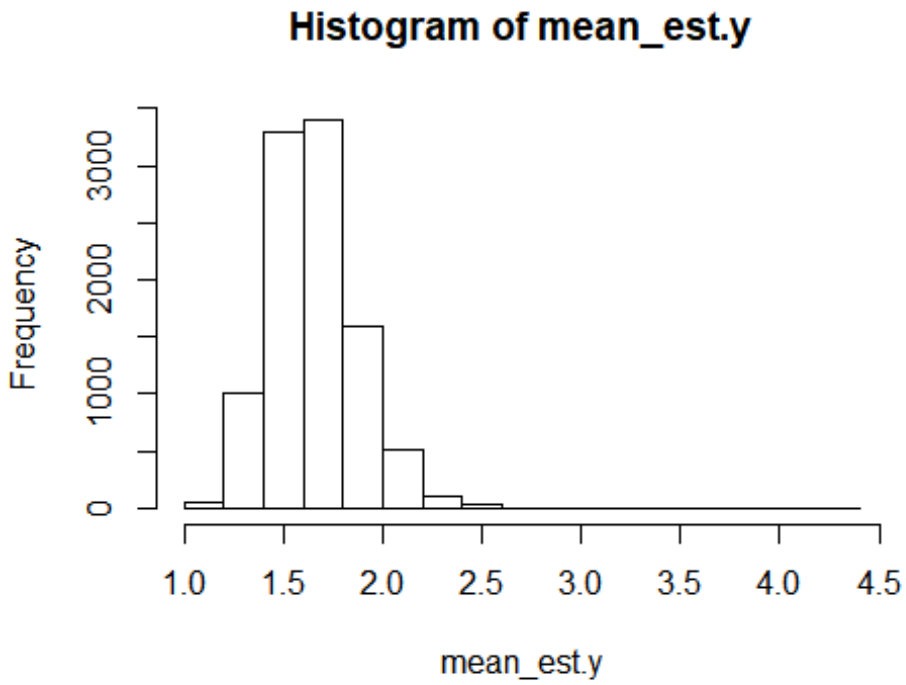
```r
set.seed(12345)
n = 100
n.sims <- 10000

est.b0 <- numeric(n.sims)
mean_est.b0 <- numeric(n.sims)
median_est.b0 <- numeric(n.sims)
mean_est.y <- numeric(n.sims)
median_est.y <- numeric(n.sims)

ci.b0 <- matrix(0, nrow = n.sims, ncol = 2)
mean_ci.b0 <- matrix(0, nrow = n.sims, ncol = 2)
median_ci.b0 <- matrix(0, nrow = n.sims, ncol = 2)
ci.mean_y <- matrix(0, nrow = n.sims, ncol = 2)
ci.median_y <- matrix(0, nrow = n.sims, ncol = 2)

for (i in 1:n.sims) {
  y = exp(rnorm(n, mean = 0, sd = 1))
  fit <- lm(log(y) ~ 1)
  mean_est.y[i] <- mean(y)
  median_est.y[i] <- median(y)
  ci.median_y[i, ] <- exp(quantile(log(y), prob = c(0.025, 0.975)))

  ci.b0[i,] <- confint(fit)[1,]
}

head(ci.b0)
```

```
##             [,1]      [,2]
## [1,]  0.02401042 0.4663840
## [2,] -0.15541897 0.2458852
## [3,] -0.23120888 0.1387857
## [4,]  0.02267787 0.4078739
## [5,] -0.22302132 0.1286398
## [6,] -0.15194112 0.2802611
```

```r
hist(median_est.y)
```




```
hist(mean_est.y)
```



```r
median_ci.captured_1 <- (1 >= exp(ci.b0[,1])) & (1 <= exp(ci.b0[,2]))
table(median_ci.captured_1)
```

```
## median_ci.captured_1
## FALSE  TRUE
##   480  9520
```

```r
mean(median_ci.captured_1)
```

```
## [1] 0.952
```

```r
mean_ci.captured_2 <- (exp(1/2) >= exp(ci.b0[,1])) & (exp(1/2) <=
exp(ci.b0[,2]))
table(mean_ci.captured_2)
```

```
## mean_ci.captured_2
## FALSE  TRUE
##  9986    14
```

```r
mean(mean_ci.captured_2)
```

```
## [1] 0.0014
```

(i) Comment on the distribution of the median values - i.e. comment on central tendency, spread, and shape

The histogram of median value shows a central tendecy at around 1, spread is from around 0.6 to 1.55, and has a bell shape.

(ii) Repeat (i) for the mean values.

The histogram of mean value shows a central tendency at around 1.6, spread is from around 1 to 2.8, and is skewed to the right.

(iii) How many of the resulting back transformed confidence intervals contain the true median value of exp(0) = 1

There were 9447 back transformed confidence intervals contain the true median value of 1.

(iv) How many of the resulting back transformed confidence intervals contain the true mean value of exp(1/2) = 1.649

There were 14 back transformed confidence intervals containing the true mean value of exp(1/2)

(v) Based on your answers to (iii) and (iv) are the back transformed intervals relevant for estimating the mean or for estimating the median? Explain your answer.

The back transformed intervals are relevant for estimating the median as there were 9520 back transformed confidence intervals containing the true median value in comparison with 14 back transformed confidence intervals containing the true mean value of exp(1/2).

Question 2

Create a single sample of 100 observation from the log-Normal distribution (as you did in the previous question). Use the non-parametric bootstrap to find a 95% CI for the median and a 95% CI for the mean of the log-Normal distribution. Do these intervals capture the true values of the median and mean respectively?

```
set.seed(12345)
n = 100
n.sims <- 10000
np.est.b0 <- numeric(n.sims)
mean_y <- numeric(n.sims)
median_y <- numeric(n.sims)

y <- exp(rnorm(n, mean = 0, sd = 1))

for (i in 1:n.sims) {
  samp = sample(1:n, replace = T)
  boot <- y[samp]
  mean_y[i] <- mean(boot)
  median_y[i] <- median(boot)

}
head(mean_y)

## [1] 2.346311 2.645691 2.193522 2.333682 2.279874 2.353445

head(median_y)

## [1] 1.738988 1.765364 1.421429 1.655501 1.225441 1.634257

quantile(median_y, prob=c(0.025, 0.975))

##      2.5%     97.5%
## 0.9757553 1.8443437

quantile(mean_y, prob=c(0.025, 0.975))
```

```
##     2.5%    97.5% 
## 1.761557 2.725540
```

The intervals capture the true values of the median and mean respectively.

Question 3

(i) Create a data frame containing this data. Designate pip as a factor and print

out the data set.

```
df = data.frame("pyr" = c(1.50, 1.06, 0.75, 1.10, 0.78, 0.55, 0.80, 0.57, 0.40, 0.65, 0.46, 0.32, 0),
           "pip" = c("0", "0", "0", "0.25", "0.25", "0.25", "2.5", "2.5", "2.5", "10", "10", "10", "0"),
           "n" = c(150, 149, 150, 151, 151, 150, 149, 150, 140, 150, 150, 149, 200),
           "y" = c(138, 75, 32, 129, 65, 19, 143, 112, 37, 141, 117, 56, 1))
df
```

```
##     pyr  pip   n   y
## 1  1.50    0 150 138
## 2  1.06    0 149  75
## 3  0.75    0 150  32
## 4  1.10 0.25 151 129
## 5  0.78 0.25 151  65
## 6  0.55 0.25 150  19
## 7  0.80  2.5 149 143
## 8  0.57  2.5 150 112
## 9  0.40  2.5 140  37
## 10 0.65   10 150 141
## 11 0.46   10 150 117
## 12 0.32   10 149  56
## 13 0.00    0 200   1
```

(ii) Fit a logistic regression model that uses both pyr and pip to model the probability of mortality. For this model treat pip as a factor and pyr as a continuous numeric variable. Do the appropriate added variable test to see if there is evidence that the interaction between pip and pyr should be included in the model. What do you conclude from this test?

```
int_fit <- glm(cbind(y, n-y) ~ pyr * pip, family = binomial, data = df)
simple_fit <- glm(cbind(y, n-y) ~ pyr + pip, family = binomial, data = df)
anova(int_fit, simple_fit, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(y, n - y) ~ pyr * pip
## Model 2: cbind(y, n - y) ~ pyr + pip
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1         5      4.626
## 2         8     50.818 -3  -46.192 5.162e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

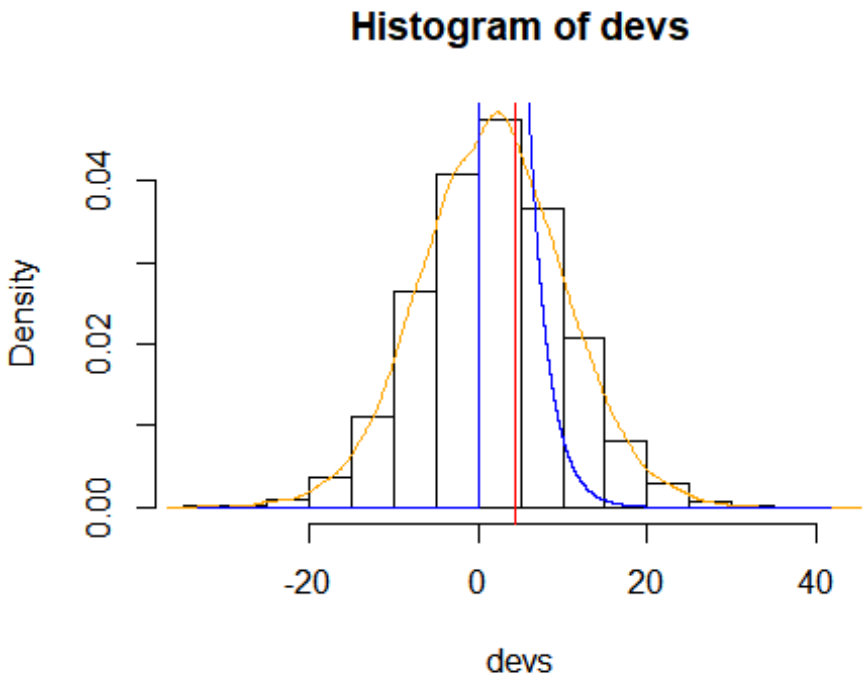
We can observe that there is evidence that the interaction between pip and pyr should be included in the model. All the interaction terms, pyr:pip0.25, pyr:pip10, and pyr:pip2.5 are statistically signficiant. Moreover, by comparing deviances via anova funciton, we found evidence that the submodel is not appropriate, our model needs at least one of the interaction terms. We can conclude that we found evidence that mixtures of pyrethrins and piperonyl have toxicity effect to red flour beetles.

(iii) Consider the test in (ii). What is used as the test statistic and what is used as the reference distribution? Use the parametric bootstrap to generate an empirical sampling distribution for this test statistic. Compare your empirical sampling distribution to the (theory based) reference distribution. Does this affect your conclusion about whether or not the interaction should be included in the model?

Chi squared statistic was the test statistc and the reference distribution was chi squared distribution. The comparison with empirical sampling distribution to the reference distribution doesn't affect my conclusion about whether the interaction should be included in the model. The sampling distribution of the residual deviance is well approximated by chi squared distribution with df 4.28.

```
set.seed(1235)
N.sim <- 10000
n.obs <- nrow(df)
```

```r
n = 100

# make a vector
pyr_pip0.25 = df$pyr[4:6]
pyr_pip2.5 = df$pyr[7:9]
pyr_pip10 = df$pyr[10:12]

devs = numeric(N.sim)
simp_cfs = coef(simple_fit)
cfs = coef(int_fit)

simple_first_xbeta_pip0 = simp_cfs[1] + simp_cfs[2]*df$pyr[1:3]
simple_xbeta0.25 <- simp_cfs[1] + simp_cfs[2]*pyr_pip0.25 + simp_cfs[3]
simple_xbeta2.5 <- simp_cfs[1] + simp_cfs[2]*pyr_pip2.5 + simp_cfs[5]
simple_xbeta10 <- simp_cfs[1] + simp_cfs[2]*pyr_pip10 + simp_cfs[4]
simple_last_xbeta_pip0 = simp_cfs[1] + simp_cfs[2]*df$pyr[13]
simple_xbeta <- c(simple_first_xbeta_pip0, simple_xbeta0.25, simple_xbeta2.5,
simple_xbeta10, simple_last_xbeta_pip0)

first_xbeta_pip0 = cfs[1] + cfs[2]*df$pyr[1:3]
last_xbeta_pip0 = cfs[1] + cfs[2]*df$pyr[13]
xbeta_pip0.25 = cfs[1] + cfs[2]*pyr_pip0.25 + cfs[3] + cfs[6] * pyr_pip0.25
xbeta_pip2.5 = cfs[1] + cfs[2]*pyr_pip2.5 + cfs[5] + cfs[8] * pyr_pip2.5
xbeta_pip10 = cfs[1] + cfs[2]*pyr_pip10 + cfs[4] + cfs[7] * pyr_pip10
xbeta = c(first_xbeta_pip0, xbeta_pip0.25, xbeta_pip2.5, xbeta_pip10,
last_xbeta_pip0)

for (i in 1:N.sim) {
  ysim_simple <- rbinom(n.obs, size = n, prob =
exp(simple_xbeta)/(1+exp(simple_xbeta)))
  ysim_int <- rbinom(n.obs, size=n, prob = exp(xbeta)/(1+exp(xbeta)))
  mod_i <- glm(cbind(ysim_int, n-ysim_int) ~ pyr * pip, family = "binomial",
data = df)
  simple_mod_i <- glm(cbind(ysim_simple, n-ysim_simple) ~ pyr + pip, family =
"binomial", data = df)
  devs[i] <- deviance(simple_mod_i) - deviance(mod_i)
}

hist(devs, freq = FALSE)
lines(density(devs), col = "orange")
ds = seq(min(devs), max(devs), length = 1e4)
lines(ds, dchisq(ds, df = (df.residual(simple_mod_i) - df.residual(mod_i))),
col = "blue")
abline(v = deviance(simple_mod_i) - deviance(mod_i), col = "red")
```




```
deviance(simple_mod_i) - deviance(mod_i)
```

```
## [1] 4.276614
```

(iv) Use the model you selected to estimate the concentration of pyrethrin needed to have a .80 probability of mortality in flour beetles when no piperonyl butoxide is used. Repeat this calculation for solutions that contain 10% piperonyl butoxide.

```
# make a vector
coef_int <- coef(int_fit)
(log(4) - coef_int[1])/(coef_int[2])
```

```
## (Intercept)
##    1.314023
```

```
(log(4) - coef_int[1] - coef_int[4])/(coef_int[2] + coef_int[7])
```

```
## (Intercept)
##   0.4906598
```

Using the model with interaction term, the estimated concentration of pyrethrin needed to have a 0.8 probability of mortality in flour beetles when no puperonyl butoxide is used is 1.31%. Whereas, the estimated concentration of pyrethrin needed for 10% piperonyl butoxide for 0.8 mortality in flour beetles is 0.49%.

(v) Use the parametric bootstrap to create 95% confidence intervals for each of your estimates from (iv).

```
set.seed(12345)
betas = matrix(0, nr=N.sim, nc= 8)

for (i in 1:N.sim) {
  ysim = rbinom(n.obs, size=n, prob = exp(xbeta)/(1+exp(xbeta)))
  mod_i = glm(cbind(ysim, n-ysim) ~ pyr * pip, family = binomial, data = df)
  coef(mod_i)
  betas[i,] = coef(mod_i)
}

quantile((log(4)-betas[,1])/betas[,2], c(0.025, 0.975))

##     2.5%    97.5%
## 1.680457 1.817198

quantile(((log(4)-betas[,1]-betas[,4])/(betas[,2] + betas[,7])), c(0.025,
0.975))

##      2.5%     97.5%
## 0.7622739 0.8759655
```

The estimated concentration of pyrethrin needed to have 0.8 probability of mortality in flour beetles when no piperonyl butoxide is somewhere between 1.25% and 1.38%. The estimated concentration of pyrethrin needed to have 0.8 probability of mortality in flour beetles when the solution contains 10% piperonyl butoxide is somewhere between 0.46% and 0.53%.

(vi) Based on these results, briefly comment on the effectiveness of piperonyl butoxide as a synergist.

Based on the results we've received, piperonyl butoxide is a good synergist with pyrethrin to kill flour beetles. According to our estimation, we only needed 0.5% of pyrethrin when the solultion had 10% piperonyl butoxide. The parametric bootstrap supports the claim as the solution without piperonyl required 1.23% to 1.4% of pyrethrin have the same effectiveness with the solution with piperonyl requiring only 0.46% to 0.53% of pytherin.