

## Stat 416 Homework 4 Solutions

Exam Corrections (10 points)

1. (6 points) For the complete experiment, we have the following table of factors with levels.

Factor	Levels
Diet	chimp McDonald's cafe mouse
Tissue	brain liver
Litter	levels were not specified in the paper
Batch	batch 1 batch 2

For the experimental data and complete information that is available to us, we have the following table of factors with levels.

Factor	Levels
Diet	chimp McDonald's cafe mouse
Batch	batch 1 batch 2

Note that tissue is not a factor in our data set because it is constant (only liver data are available). From reading the paper, it appears that the researchers tried to treat litters as blocks, but we don't know how many litters there were or how many experimental units were in each litter or how treatments were assigned within litter.

2. (6 points) In the full experiment, mice are whole-plot experimental units, and each sample (brain or liver) are like split-plot experimental units. In our data, we have only mice as experimental units.
3. (6 points) The complete experiment was some form of a split-plot experiment. For the data that we have, I would treat it as a randomized complete block design with replication in each block. The blocks are batches. The treatment of interest is diet. The paper did not specify how randomization played a role in the design. The experiment probably involved cages, though the authors make no mention of this factor. If the mice were individually caged, the cage factor is not important, but if mice were grouped in cages, this should be accounted for in the analysis.
4. (6 points) The litter factor was mentioned by the authors, but there is no information in the data set to tell use which mice were from which litters.
5. `d=read.table("GSE6297_series_matrix.txt", skip=67, nrow=45101, header=T)`

6. (6 points)

```
diet=gl(4, 6, labels=c("chimp", "McD", "cafe", "mouse"))
batch=as.factor(rep(c(1, 1, 1, 2, 2, 2), 4))
geneNames=d[, 1]
```

```
d=d[,-1]
row.names(d)=geneNames
names(d)=paste(diet,"B",batch,"R",rep(1:3,8),sep=" ")
boxplot(d,col=as.numeric(batch)+1,las=3)
```

The boxplots show quite clearly that there is a substantial difference between batches. The interquartile ranges tend to be larger for batch 1 chips than for batch 2 chips.

7. (6 points) The paper states that RMA was used to compute the expression measures.

8. (a) (6 points)

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
as.factor(a)	1	0.0040	0.0040	0.0426	0.8409932
as.factor(b)	1	3.2033	3.2033	33.8685	0.0002532 ***
Residuals	9	0.8512	0.0946		

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The p-values for factor a is about 0.8409932. The p-value for factor b is 0.0002532

(b) (6 points) For example,

$$\mu = 100, \alpha_1 = -94.8, \alpha_2 = -98.9, \beta_1 = 0, \beta_2 = 2.3$$

(c) (6 points)

```
> co=coef(out)
> co
(Intercept) as.factor(a)2 as.factor(b)2
6.03500000 -0.03666667 -1.03333333
> matrix(c(co[1],co[1]+co[2],co[1]+co[3],sum(co)),nrow=2)
      [,1]      [,2]
[1,] 6.035000 5.001667
[2,] 5.998333 4.965000
```

(d) (6 points)

$$(\mu + \alpha_1 + \beta_1) - (\mu + \alpha_2 + \beta_2) = \alpha_1 + \beta_1 - \alpha_2 - \beta_2$$

which is estimated by R as  $-\hat{\alpha}_2 - \hat{\beta}_2$ .

```

> co=coef(out)
> co
(Intercept) as.factor(a)2 as.factor(b)2
6.03500000 -0.03666667 -1.03333333

> -co[2]-co[3]
as.factor(a)2
1.07

```

-1.07 is also an acceptable answer because the direction of subtraction was not specified.

(e) (10 points)

```

> a=anova(out)
> df=a[3,1]
> b=coef(out)
> v=vcov(out)
> m=c(0,-1,-1)
> tstat=t(m)%*%b/sqrt(t(m)%*%v%*%m)
> p=2*(1-pt(abs(tstat),df))
> p
[1,]
[1,] 0.002107592

```

The small p-value suggests that the mean difference is significantly different from 0.

9. (a) (14 points) The following function provides a p-value for the test of diet effects, the test of batch effects, the test for a difference between the chimp diet and the McDonald's diet, and the test for a difference between the cafe diet and the mouse diet. Other tests can be obtained in an analogous manner.

```

get.p=function(y)
{
  out=lm(y~diet+batch)
  a=anova(out)
  df=a[3,1]
  b=coef(out)
  v=vcov(out)
  m=c(0,1,0,0,0)
  tstat=t(m)%*%b/sqrt(t(m)%*%v%*%m)
  p12=2*(1-pt(abs(tstat),df))
  m=c(0,0,-1,1,0)

```

```

tstat=t(m)**b/sqrt(t(m)**v**m)
p34=2*(1-pt(abs(tstat),df))
p=c(a[1:2,5],p12,p34)
p
}

get.p(as.numeric(d[1,]))
[1] 1.885223e-03 2.870306e-06 8.388906e-01 8.121337e-04

```

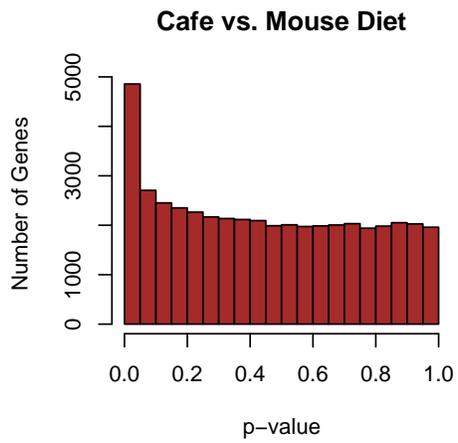
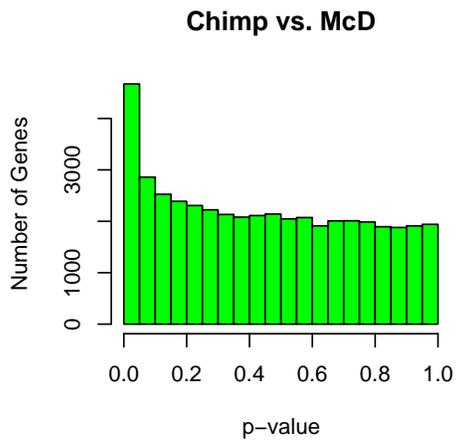
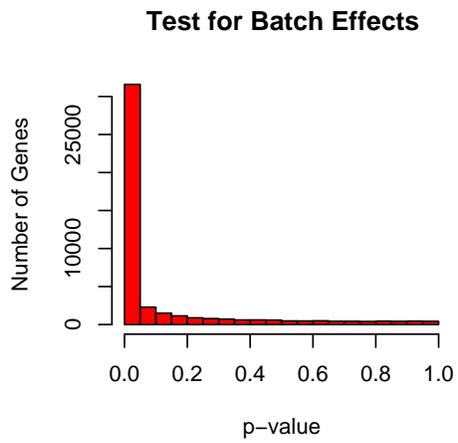
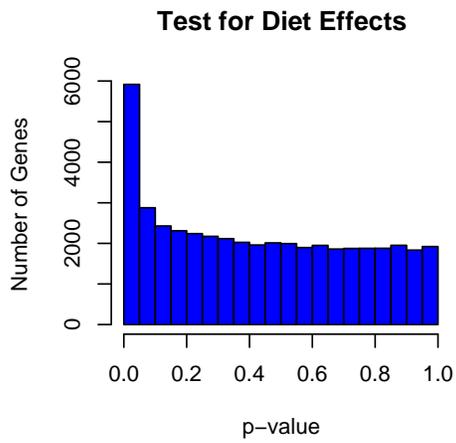
(b) (6 points)

```

p=t(apply(d,1,get.p))

par(mfrow=c(2,2))
hist(p[,1],xlab="p-value",ylab="Number of Genes",
     main="Test for Diet Effects",col="blue")
hist(p[,2],xlab="p-value",ylab="Number of Genes",
     main="Test for Batch Effects",col="red")
hist(p[,3],xlab="p-value",ylab="Number of Genes",
     main="Chimp vs. McD",col="green")
hist(p[,4],xlab="p-value",ylab="Number of Genes",
     main="Cafe vs. Mouse Diet",col="brown")

```



The p-value distributions show evidence of differences among diets and extreme differences due to batch effects. The vast majority of genes had different expression levels across the two batches.