Statistics with R for Biologists

James H. Bullard
Kasper Daniel Hansen
Margaret Taub

Berkeley, California
July 7-11, 2008

Case Study: Smith et. al. Gene-Environment Interaction in Yeast Gene Expression

Background

- In Smith et. al. the authors wish to assess the effects of yeast strain (gene) and condition (environment) on the phenotype gene expression.
- The authors have hybridized: 2 (strain) * 2 (condition) * 2 (dye). The have replicated this 3 times for a total of 24 hybridizations.
- All hybridizations were done using two-color 11k Agilent yeast arrays. All samples were hybridized against a common reference sample.
- Data was pre-processed using Agilent software to perform quality control (outlier removal) leaving a total of 4,342 "high-quality" transcripts for the "parental analysis."

"Parental Analysis"

We will focus exclusively on the first portion of their analysis. The question they wish to answer is: what genes show significant strain-condition interaction? They want to determine which genes are better described by the model:

$$\text{phenotype} \sim \text{dye} + \text{strain} + \text{condition} + \text{strain} * \text{condition}$$

As compared to:

$$\text{phenotype} \sim \text{dye} + \text{strain} + \text{condition}$$

Getting Started

Example

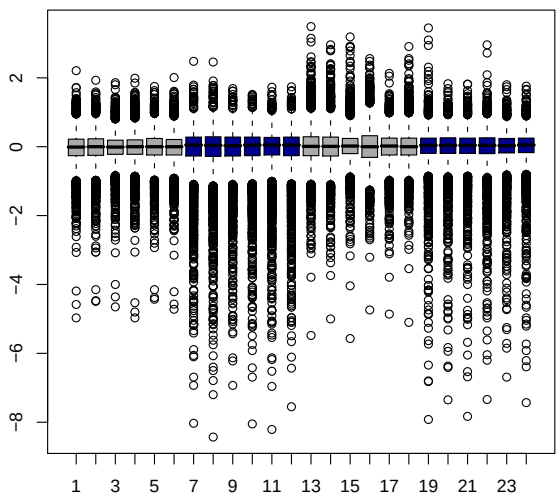
We first want to read in the data and convince ourselves that we have the same data that they have used to conduct their analysis. I have provided two .csv files from the paper to use in reproducing/understanding their analysis. The files are: smith_et_al_data.csv, smith_et_al_pvals.csv. First, read in smith_et_al_data.csv and have a look at the data. We want a matrix with the row names equal to the gene names and the columns a different factor-level combination. For fun do a t-test comparing the means for each gene within a particular condition – which gene has the largest t-statistic? what does this mean? How many numbers contribute to each t-statistic? How many are "significant"?

T-tests

	glucose	ethanol
FALSE	1189	1963
TRUE	3153	2379

Table 1: Number of genes reported as differentially expressed between strains at the .05 % cutoff.

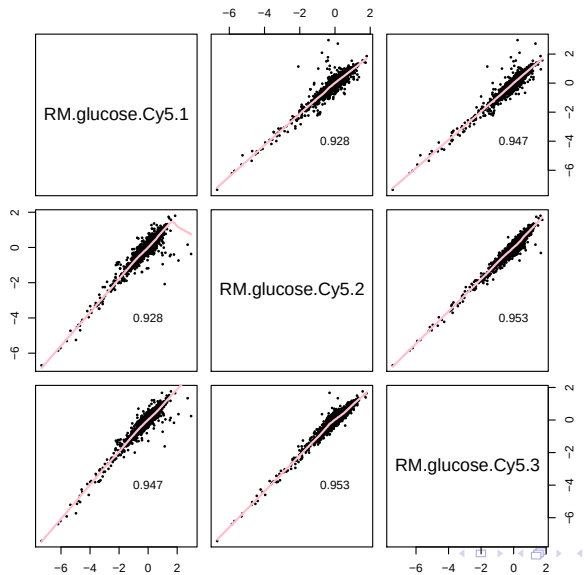
T-tests



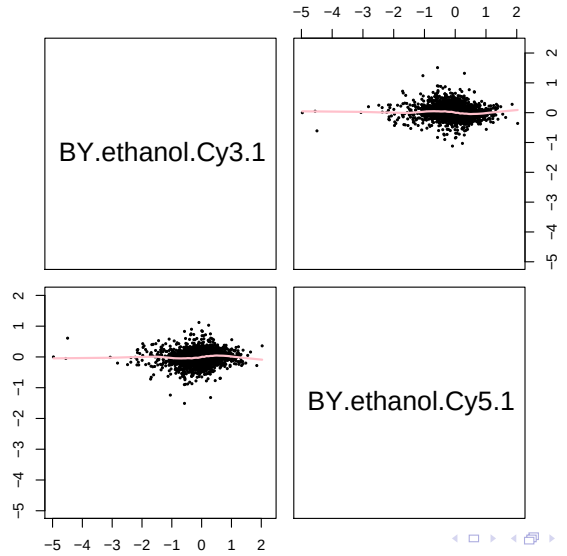
Visualization

1. Create a pairs plot for each of the 8 sets of three replicates.
2. Create mean difference plots comparing the replicate experiments as well as dye-swaps.
3. Create image plots of the microarray data sets.

Visualization



Mean-Difference Plots



Experimental Design

- In this experiment we are primarily interested in the effects of strain and condition on phenotype (gene-expression).
- In addition, we have to worry about the effect of dye on transcription (why?).
- This experiment is an example of a complete balanced design, where each factor-level occurs as at least once.
- We might imagine a simpler experiment where we have only 8 or 16 microarrays and 1 and 2 less replicates. What do the replicates give us?

Experimental Design

```
, , dye = Cy3

      condition
strain ethanol glucose
    BY       3       3
    RM       3       3

, , dye = Cy5

      condition
strain ethanol glucose
    BY       3       3
    RM       3       3
```

Statistical Models

- How do we "represent" this experiment as a statistical model?
- We are frequently interested in the dependence of an outcome (phenotype) on a number of predictors. In this case we are interested in the effect of the predictors (strain, condition, dye) on our phenotype and we can represent an additive dependence in the following way:

$$\text{phenotype} \sim \text{dye} + \text{strain} + \text{condition}$$

We can write this more explicitly as:

$$\text{phenotype}_j = \beta_0 + \beta_1 \text{dye}_j + \beta_2 \text{strain}_j + \beta_3 \text{condition}_j + \epsilon_j$$

Statistical Models

Here j ranges over the subjects in our experiment; the independent observational units. In our current example we have 24 expression measures (phenotype) each expression measure was obtained from an experiment conducted at a particular dye, strain, condition combination. If I tell you that dye $= 1$, strain $= 0$, and condition $= 1$, what is the phenotype?

- Here we are explicitly stating that phenotype depends on dye, strain, and condition in an additive fashion. In this model we can interpret the βs in a relatively straightforward fashion. "After fitting our model we found the value of β_2 to be equal to 4 this means that by flipping strain from 0 to 1 we can (increase | decrease) gene expression (1 | 2 | 3 | 4 | 8) times."

The Linear Regression Model

- The linear regression model is one of the most common, if not, the most common way of modeling data.
- In many cases the model is not "correct" but is often very reasonable.

$$Y = X\beta + \epsilon \tag{1}$$

The linear regression model is composed of an $n \times p$ design matrix (X), an $n \times 1$ vector of outcomes (Y), a $p \times 1$ vector of parameters which we wish to estimate (generally denoted $\hat{\beta}$). Linear regression finds the estimate $\hat{\beta}$ which minimizes the L_2 loss (equation: (2)).

$$L_2(\beta) = \sum_{i=1}^{n} (Y - X\beta)^2 \tag{2}$$

The Linear Regression Model

Under the following assumptions linear regression is the best linear unbiased estimator of β.

i. X and Y satisfy equation (1).
ii. The disturbance terms ϵ_i are i.i.d with mean 0 and variance σ^2.
iii. X and ϵ are independent.

Linear Models

- Statistical models in R have a special syntax (the formula syntax):

```
Y ~ X
```

- This says that the variable Y is related to X. The formula specification is used in a variety of functions as input and depending on that function different relationships between the predictor variables (X) and the outcome variables (Y) are assumed.

Formulas Continued

The simplest data set to begin to play with the formula functions in R can be generated as follows:

```
> N <- 100
> X <- runif(N, 20, 40)
> Y <- 3 + X * 2 + rnorm(N, mean = 0,
+     sd = 5)
```

Now suppose we would like to fit a linear model to the data. In R this is as simple as:

```
> lm.1 <- lm(Y ~ X)
> lm.1.int <- lm(Y ~ 1 + X)
```

1. What is the class of `lm.1` and `lm.1.int`?
2. How can we extract the estimates $\hat{\beta}$?
3. What are the functions which are specialized for this class (hint methods)?

Formulas

As the above model is not that interesting we might be inclined to to have a look at some more interesting data sets. Let's have another look at our viral load data set.

```
> vL <- read.table("../../data/viral-load.dta")
> lm.vL <- lm(viral.load ~ age +
+     meds + infected, data = vL)
```

1. Is this a sensible thing to do?
2. What are the estimates of the coefficients?
3. What happened with meds?
4. How do we transform the data to get on safer ground? (hint: try to put the log directly in the formula), try to square the age covariate.

Back to The Smith et. al. Dataset

$$\text{phenotype}_j = \beta_0 + \beta_1 \text{dye}_j + \beta_2 \text{strain}_j + \beta_3 \text{condition}_j + \epsilon_j$$

- What do each of the β coefficients represent?
- What kind of variables are dye, condition, and strain?

	strain	condition	dye
1	BY	ethanol	Cy3
4	BY	ethanol	Cy5
7	BY	glucose	Cy3
10	BY	glucose	Cy5
13	RM	ethanol	Cy3
16	RM	ethanol	Cy5
19	RM	glucose	Cy3
22	RM	glucose	Cy5

Back to The Smith et. al. Dataset

- Parameterization of the model can be quite tricky. Here we need to understand what happens with the factors dye, strain, and condition in the formula in order to fully appreciate what the βs represent. We just want to skim the surface here.
- Does this code work:

```
> genotype <- sample(c("AA", "AB",
+     "BB"), size = 100, replace = TRUE)
> cholesterol <- 160 + 3 * genotype +
+     rnorm(100)
```

- So what I really need to do is the following:

Back to The Smith et. al. Dataset

```
> genotype <- sample(c("AA", "AB",
+     "BB"), size = 100, replace = TRUE)
> genotype <- factor(genotype)
> designMatrix <- model.matrix(~genotype,
+     data = genotype)
> head(designMatrix)
```

```
  (Intercept) genotypeAB genotypeBB
1           1          0          0
2           1          1          0
3           1          0          1
4           1          0          0
5           1          1          0
6           1          0          1
```


Back to The Smith et. al. Dataset

- What we have done is convert the "factors" into "dummy" variables so that we can do some matrix algebra on them. What happened to genotype AA?
- Now we can simulate some data quite simply:

```
> cholesterol <- designMatrix %*%
+     c(160, -40, -20) + rnorm(100,
+     sd = 10)
```

- In summary, when we have factors we code them as dummy variables and we drop one of the levels – this level becomes the baseline which we compare the resulting coefficients against. In the example above having genotype BB makes your cholesterol how much higher than having genotype AA?

Fitting the Model

- The next step is that we want to "fit" the model.
- Again, we fit the model using least squares.

```
> lm(cholesterol ~ genotype)

Call:
lm(formula = cholesterol ~ genotype)

Coefficients:
(Intercept)   genotypeAB   genotypeBB
     158.68       -40.71       -17.29

> lm(cholesterol ~ genotype - 1)
```

Fitting the Model

```
Call:
lm(formula = cholesterol ~ genotype - 1)

Coefficients:
genotypeAA  genotypeAB  genotypeBB
     158.7       118.0       141.4
```


Causation

- Causation is a tricky subject. When we perform an experiment where we vary the levels deliberately we often think that the thing we vary is "causing" the change in outcome.
- In our simple cholesterol example we can see that in fact genotype does cause an increase in cholesterol – we simulated the data so we say what happens. However you can imagine receieving the data set with cholesterol and bodyFat below.
- Does "bodyFat" cause cholesterol? How would you know just by fitting the model? What experiment could you conduct to get to the bottom of causation?

```
> head(dta)
```


Causation

```
  cholesterol    bodyFat
1    138.9968 0.01048154
2    130.7072 0.22098173
3    134.0498 0.54748434
4    157.0508 0.09561934
5    109.4438 0.28014284
6    144.5649 0.52458144
```

```
> round(coefficients(summary(lm(cholesterol ~
+     bodyFat))), 4)
```


Causation

```
            Estimate Std. Error t value
(Intercept) 146.1098     3.0429 48.0169
bodyFat     -25.0994     8.8105 -2.8488
            Pr(>|t|)
(Intercept)   0.0000
bodyFat       0.0053
```

Assessing Model Parameters: Significance

- The next step is to decide whether or not a parameter is a good "predictor" of our outcome. At this point we have to discuss "statistical significance."
- In our cholesterol example the two genotypes are wildly significant – what does this mean?

```
> round(coefficients(summary(lm(cholesterol ~
+     genotype))), 4)
```


Assessing Model Parameters: Significance


```
             Estimate Std. Error
(Intercept) 158.6779     1.4313
genotypeAB  -40.7090     2.0092
genotypeBB  -17.2923     2.0242
             t value Pr(>|t|)
(Intercept) 110.8616        0
genotypeAB  -20.2608        0
genotypeBB   -8.5428        0
```


Assessing Model Parameters: Simulation

Now we want to simulate data to solidify some of the concepts above. First, we need some "predictor" variables. We are going to use the design used in Smith et. al. Here we can use the function model.matrix to help us generate the data. After constructing the design matrix we are going to use this to generate some outcome variables. Use the following formula:

$$\text{phenotype} = 1 + \beta_{strain}\text{strain} + \beta_{condition}\text{condition} + \epsilon \quad (3)$$

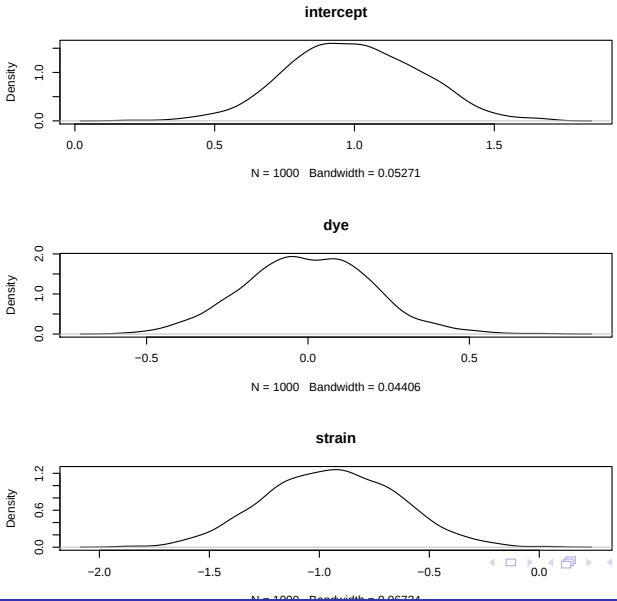
$\beta_{strain} = -.5$, $\beta_{condition} = -.95$, and $\epsilon \sim N(0, .5)$ to start. We will want to change our error distribution after we get the hang of it, but for now lets keep it simple. After we have constructed a data set, fit the model:

$$\text{phenotype} \sim \text{dye} + \text{strain} * \text{condition} \quad (4)$$

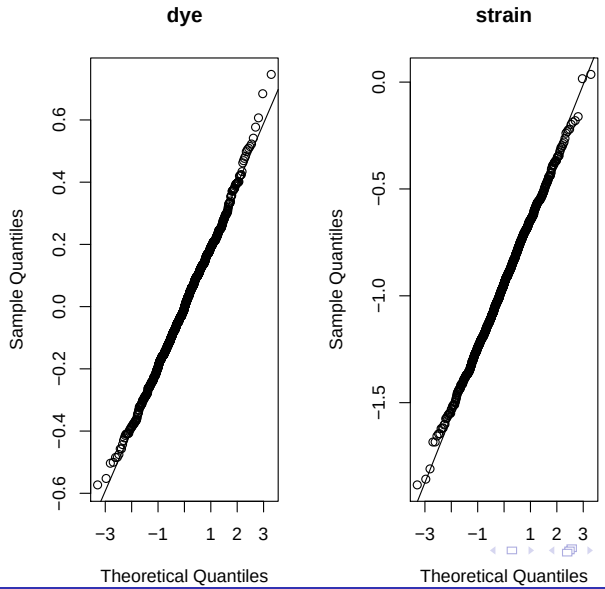
Assessing Model Parameters: Simulation

Interpret the output. After you have fit the model one time on your simulated data set we want to generate 1000 data sets and fit the model on each of these data sets. This should help us understand some of the assumptions and results of the linear model.

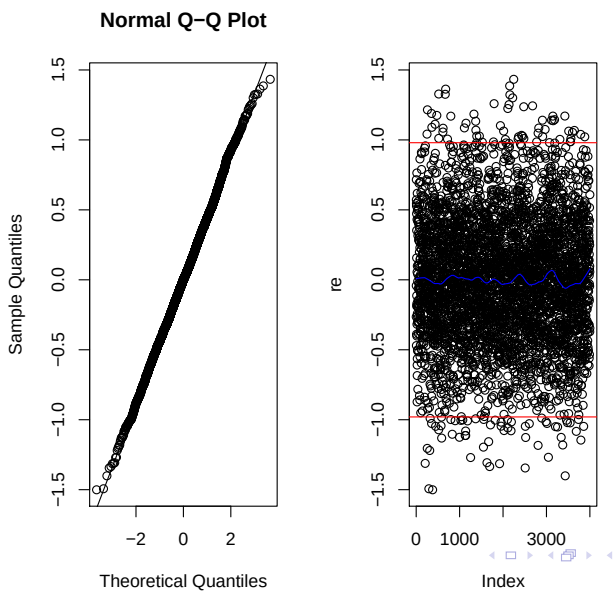
Distribution of our Estimates



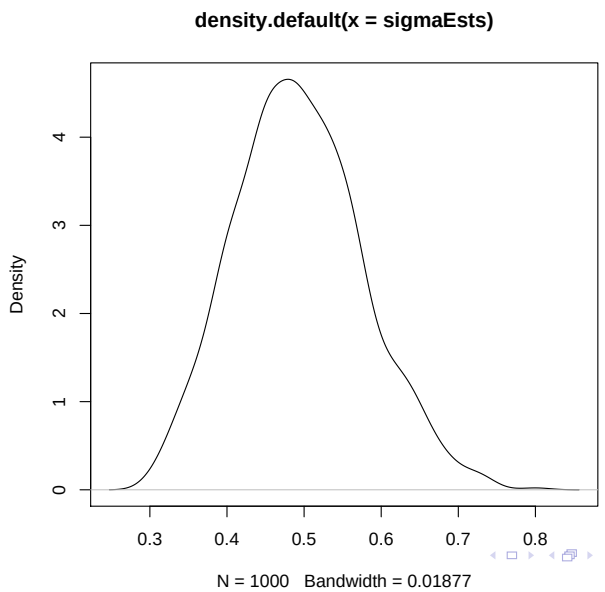
Distribution of our Estimates



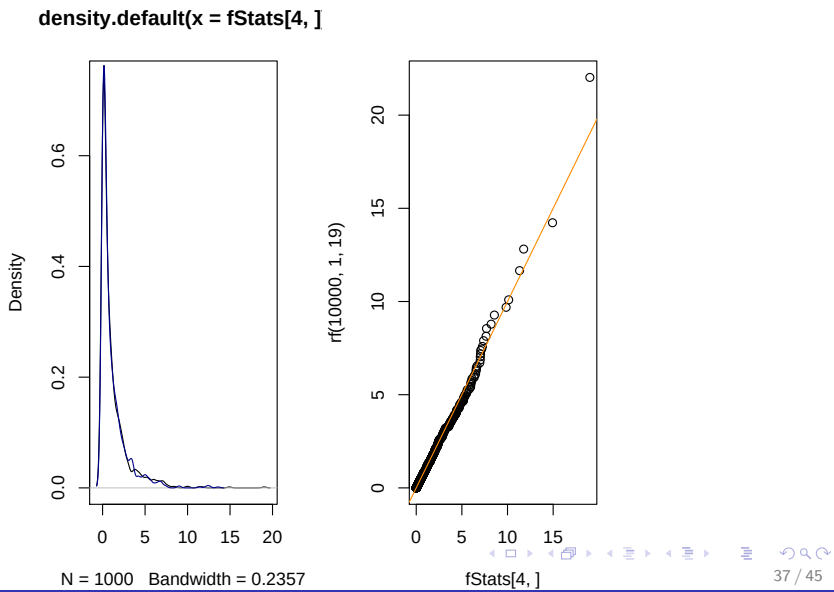
Residual Distribution



Estimating σ



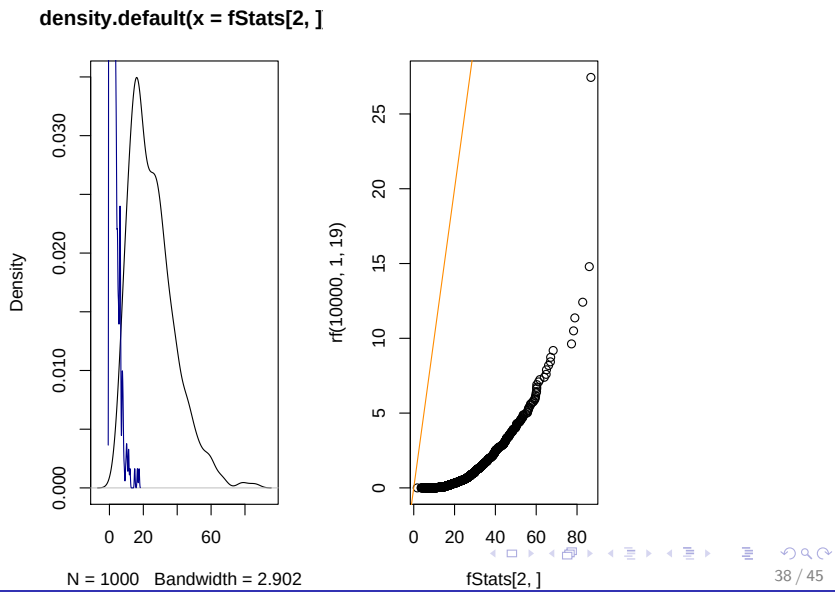
Null Distribution: F, Interaction Term



Statistics with R for Biologists

Background
Getting Started
Experimental Design
Statistical Models
Linear Models
Simulating from Smith et. al.

Null Distribution: F, Strain Term



Statistics with R for Biologists

Background
Getting Started
Experimental Design
Statistical Models
Linear Models
Simulating from Smith et. al.

Testing

- How many false positives did we commit when we looked at the F-test for the inclusion of dye in the model?

```
> sum(fStats[1, ] > qf(0.95, 1, 19))/1000
[1] 0.037
> sum(fStats[4, ] > qf(0.95, 1, 19))/1000
[1] 0.059
```

- How many false negatives did we commit?

```
> sum(fStats[2, ] < qf(0.95, 1, 19))/1000
[1] 0.006
> sum(fStats[3, ] < qf(0.95, 1, 19))/1000
[1] 0.367
```


Statistics with R for Biologists

Background
Getting Started
Experimental Design
Statistical Models
Linear Models
Simulating from Smith et. al.

Testing

Go back and change the error distribution used to simulate the 1000 data sets. Choose something with larger variance, such as a T distribution with less than 6 degrees of freedom. If you have time then go back and generate the data with a "dye" effect and then exclude that term when you fit the model.

1. What happens to our estimates?
2. What happens to the distribution of our estimates?
3. What about the distribution of our residuals?
4. What about the distribution of our test-statistic (F-statistic)?
5. What happens to the p-values, do we commit more false positives and false negatives or fewer?

ANOVA

- Analysis of Variance models are linear models with categorical predictors.
- Our last example was an ANOVA model with three factors taking on two distinct levels each. Factors can have as many discreet levels as they want, but the more levels and factors the more data you want to estimate parameters.
- In the Smith et. al. paper the statistical model which they fit is given by:

$$\text{phenotype} \sim \text{dye} + \text{strain} * \text{condition} \tag{5}$$

  Here I have written the model in terms of `R`'s notation rather than the notation in the paper.
- Is this the "full" model?

Permutation Tests

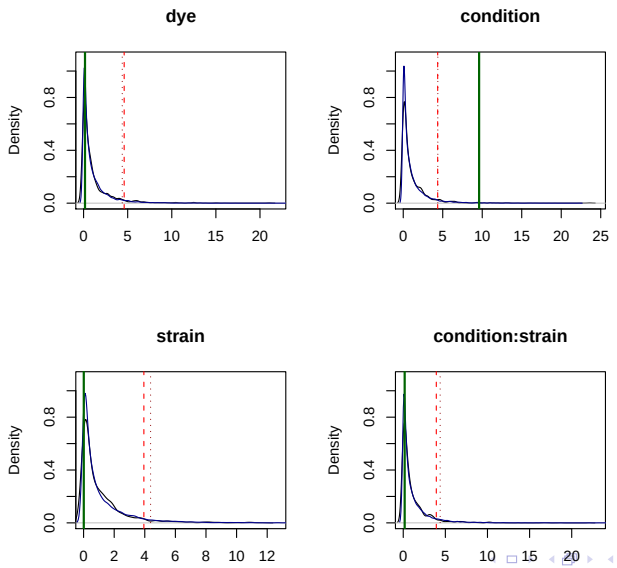
- In Smith et. al. they perform a permutation test instead of the F-test which we performed above - A permutation test allows us to construct the null distribution directly.
- As we saw above we found it relatively difficult to construct an example where the choice of an (independent) error structure induced lots of false positives.
- With a permutation test we are going to shuffle our predictors and then recompute an F-statistic, we are going to use the permutation distribution of F-statistics to test against.

Permutation Tests

Example

Using our "model.matrix" from above and normal errors simulate one data set. From this "simulated" data set construct a permutation null distribution for the F-statistics. Each F-statistic is a measure of how much evidence there is to include the term in the model as compared to the full model. Under the null distribution and some assumptions about the error distribution (ϵ is IID with normal errors and has mean 0) this F-statistic should be F distributed. After constructing a permutation distribution using the F-statistics test the observed F-statistics against this distribution.

Permutation Tests



Multiple Testing: FDR Correction

- In microarray analysis we are often testing whether a gene shows a significant deviation from some null model. One example is the null model of no differential expression.
- If I compare