Lecture 15 Lab - One-Way ANOVA

Student

4/1/2020

We have concluded our discussion of regression, which dealt with a numeric response variable we denoted y and a numeric predictor variable x or a collection of such variables.

A couple of notes:

If your response variable is categorical, like success/failure, then you need something like logistic regression, which gets covered in Math 150.

I stated above that all of our predictor variables were also numeric. We know this isn't necessary. If we had a categorical variable, we converted it into a collection of dummy variables (for which we needed one fewer than the number of "levels" the categorical variable took on).

What we are going to transition to now is where the predictor variables are all categorical. Furthermore, we are going to think of our data as arising from an experiment (rather than just observing some data). So we get to think about how we should design an experiment to get nice statistical properties, and thus improve our chances at getting valuable scientific insight.

ANOVA (The ANalysis Of VAriance)

Our question will continue to be about means μ_i. A lot of this is going to look familiar, both because the first little bit of this is standard in an introductory course, and because we have played with some of these ideas in various ways already in this course.

In regression, we assumed that the means $\mu_{y|x}$ were linear, i.e. $\beta_0 + \beta_1 x$ or some similarly "flat" higher dimensional surface if we had more than 1 x.

As our predictor is categorical, there is no such structure. We allow a different mean to be associated with every "level" of x. Wait, this sounds like the formal test of linearity and SSPE and SSLF. Yeah. It does.

What we'll start with is called Type I or fixed effects ANOVA. We'll do this for a good long while. What is the distinction?

Fixed effects means that you observe the categorical variable at every level that you could care about.

Random effects assumes that you are interested in a variable that takes on many (mathematically treated as infinitely) levels, and you only get to observe a random sample. As an example, I might wonder if the amount of salt on the fries at a national fast food restaurant is the same on average at every restaurant. We'd try to answer that question about means by only sampling a hand full of restaurants, as we couldn't hope to sample all of them. More on this in a while.

So... fixed effects. An example would go well here.

Suppose we measure the duration of a cold. We think about taking a zinc supplement, a Vitamin C supplement, both, or nothing at all. So we have 4 levels. And my response is the duration of symptoms.

My question, $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs H_a :something is different somewhere. (It doesn't work to just put inequalities everywhere, as there are possibilities that then are neither the null nor the alternative).

Let's do some classic stuff. Measure some sums of squares.

If I had to guess the duration of someone's symptoms, but didn't know which treatment they were under, I'd guess the overall average. If I knew what treatment they were under, I'd guess the treament mean. Let's get some (terrible?) notation going.

My data is y_{ij}, the j^{th} person observed under the i^{th} treatment, $i = 1, \ldots, k$ in general, with $k = 4$ in the current example.

I'll write the model as this (known as the cell means model)

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

where the same assumptions will be made (lots of similarities here).

If I knew what treatment they were in, say treatment i, I'd guess

$$\bar{y}_{i\cdot} = \frac{\sum_j y_{ij}}{n_i}$$

where n_i is the number of people in that treatment.

If $n_i = n$ for all i, then we say that we have a balanced design.

If I didn't know what treatment group they were in, I'd have to guess the overall mean

$$\bar{y}_{\cdot\cdot} = \frac{\sum_i \sum_j y_{ij}}{\sum_i n_i}$$

Note what the notation is doing. If the index is averaged over, it gets replaced with a $\cdot$ in the subscript.

Now would be a good time to get used to this notation, as it's going to get worse.

You know what I like? Sums of squares.

The person who has to guess without knowing the treatment?

$$SSTO = \sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2$$

The person who is given the treatment?

$$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2$$

The difference between these things?

$$SSTR = \sum_i n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

This can be shown via a favorite trick, adding 0 $(\bar{y}_{i\cdot} - \bar{y}_{i\cdot})$ into SSTO and expanding.

Let's get some data, and see what all of this means.

```r
mus <- c(5, 4, 4, 3)
y <- rep(mus, 10) + rnorm(40,0,1)
x <- as.factor(rep(c('Control', 'Zi', 'vC', 'ZivC'), 10))
data <- data.frame(y,x)
head(data)
```

```
##          y       x
## 1 5.512213 Control
## 2 3.606159      Zi
## 3 3.869850      vC
## 4 5.414205    ZivC
## 5 5.947629 Control
## 6 5.000396      Zi
```

So I've created the data such that if you are only going to take one supplement, it doesn't matter which ($\mu = 4$ in both cases), but taking both is helpful ($\mu = 3$).

We have a balanced design with $n_i = 10$ for all $i = 1, 2, 3, 4$.

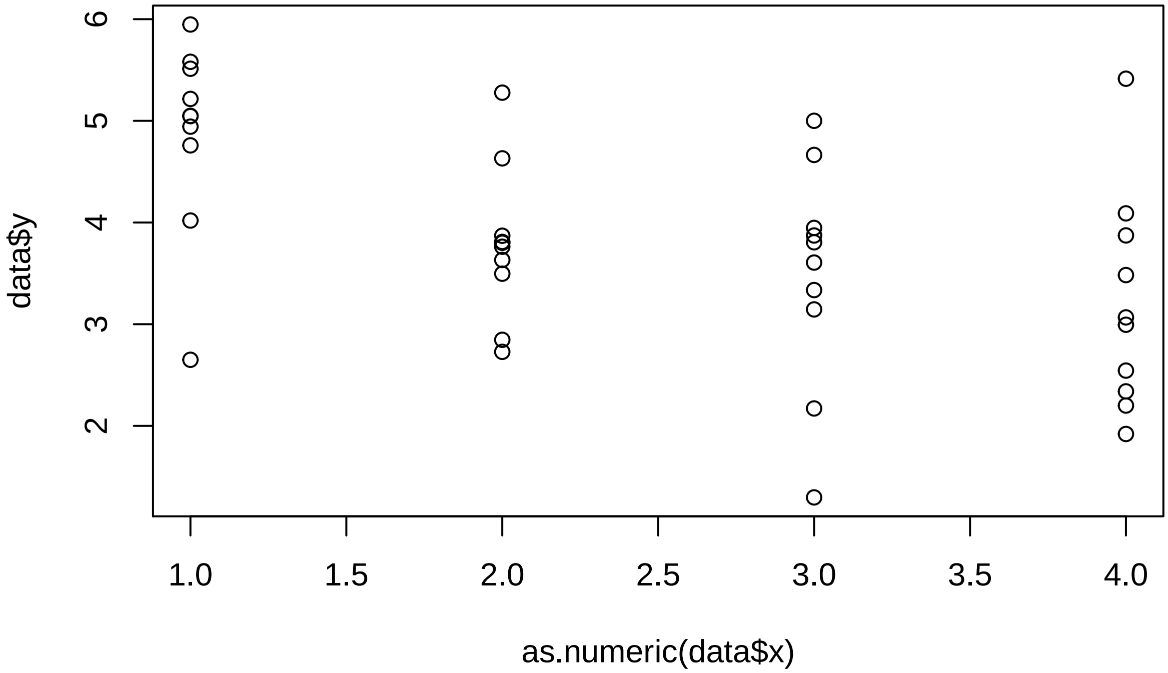
```
summary(data)
```

```
##        y                x     
##  Min.   :1.297   Control:10  
##  1st Qu.:3.049   vC     :10  
##  Median :3.807   Zi     :10  
##  Mean   :3.834   ZivC   :10  
##  3rd Qu.:4.805               
##  Max.   :5.948               
```

For instance, y_{34} is going to be the value for the 4th person assigned to the Vitamin C group. That second index is arbitrary. The first one tells us their treatment.

Let's visualize the data.

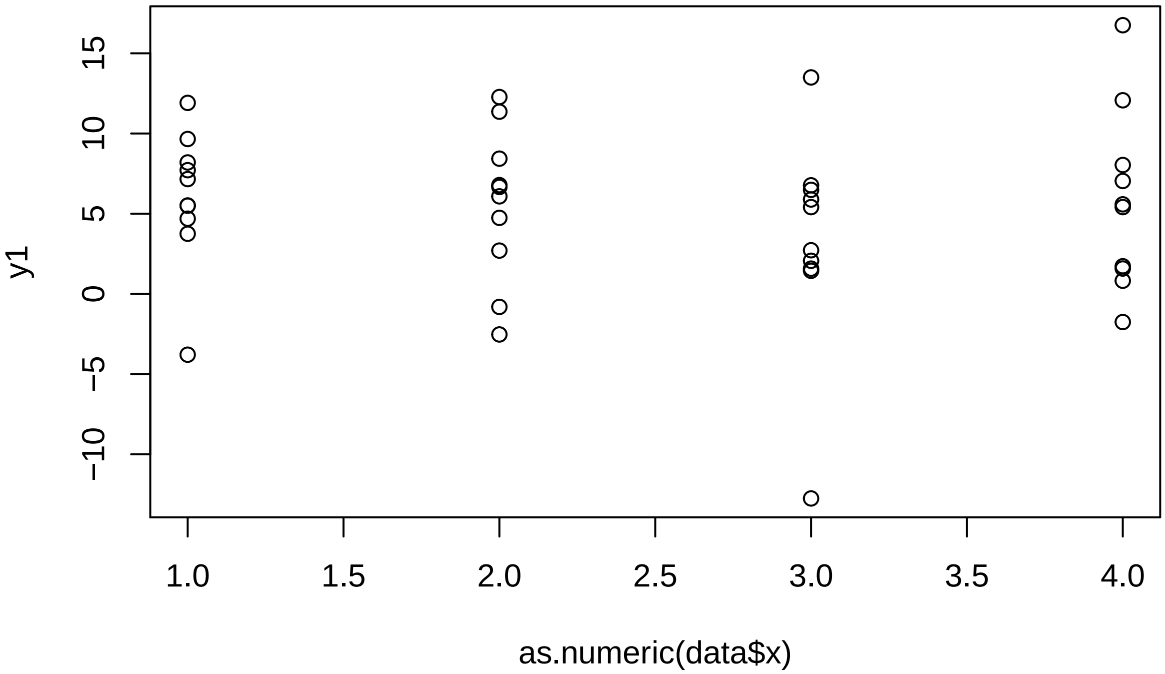
```
plot(as.numeric(data$x), data$y)
```



Are you convinced that something is going on here, that at least some of the μ_i are different from some of the other? Which ones are you convinced about (ignoring the fact that you know the truth because it's simulated)?

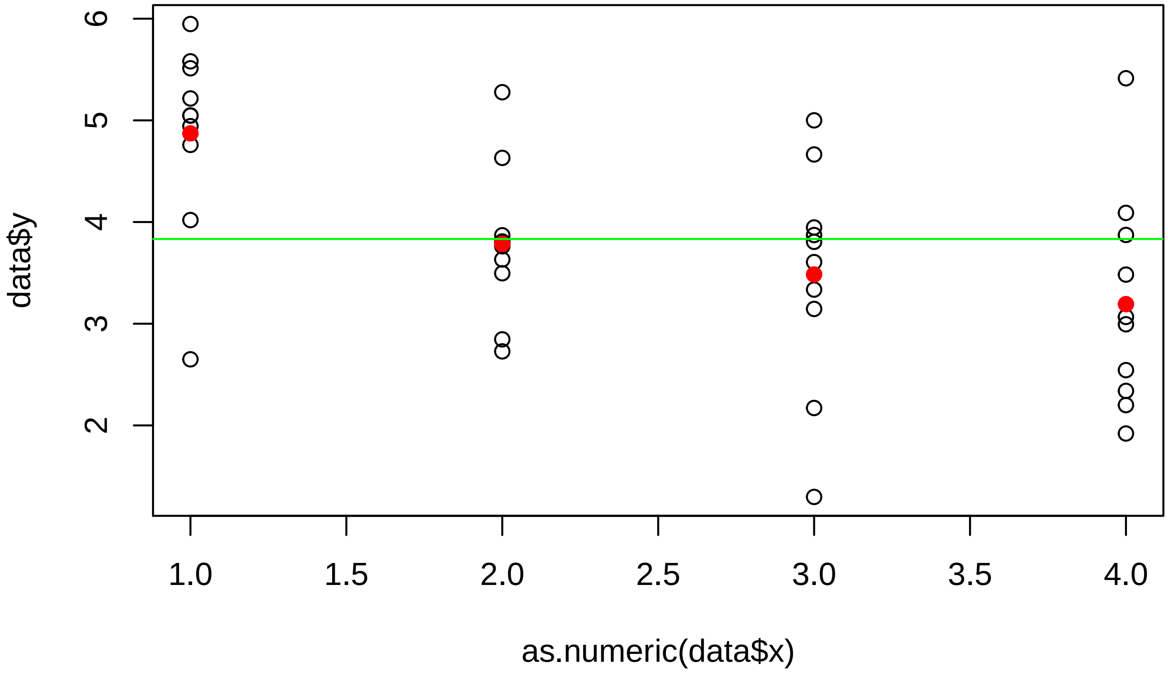
Why are you convinced? Here's a different data set, with the same μ_i values:

```
mus <- c(5, 4, 4, 3)
y1 <- rep(mus, 10) + rnorm(40,0,5)
plot(as.numeric(data$x), y1)
```



Let's go back to our original data, and we might say that we think the μ are not the same because the $\bar{y}_{i\cdot}$ are so different. Let's visualize.

```r
means <- mean(~y|x, data=data)
plot(as.numeric(data$x), data$y)
points(1:4, means, col='red', pch=19)
abline(mean(data$y), 0, col='green')
```



The red dots are the $y_{i\cdot}$ and the green line is $\bar{y}_{\cdot\cdot}$.

They are “far away” from the green line. But how to quantify that? Oh…

$$SSTR = \sum_i n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

What is this quantity? Well, let’s make things simple and assume its a balanced design. Then I can write it as

$$n\sum_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

This looks like n times the numerator of a variance formula. Let’s make it a full variance formula.

$$MSTR = \frac{SSTR}{k-1}$$

This is the variance of the $\bar{y}_{i\cdot}$, if $\bar{y}_{\cdot\cdot}$ is an estimate of the mean of $\bar{y}_{i\cdot}$ for all i, in other words, if the null is true. This is, I think, the hardest concept of the day (besides the notation).

So if the null is true, MSTR should be n times the variance of $\bar{y}_{i\cdot}$. One of the most valuable results in intro stats is that $\sigma_{\bar{y}}^2 = \frac{\sigma^2}{n}$ or $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$ as you more commonly saw it.

So if the null is true

$$MSTR \approx nVar(\bar{y}_{i\cdot}) = n\frac{\sigma^2}{n} = \sigma^2$$

If the null isn’t true, this will be bigger. You’ve seen these arguments on homeworks and the midterm. We’ll actually say what this is approximating under the alternative, but not today.

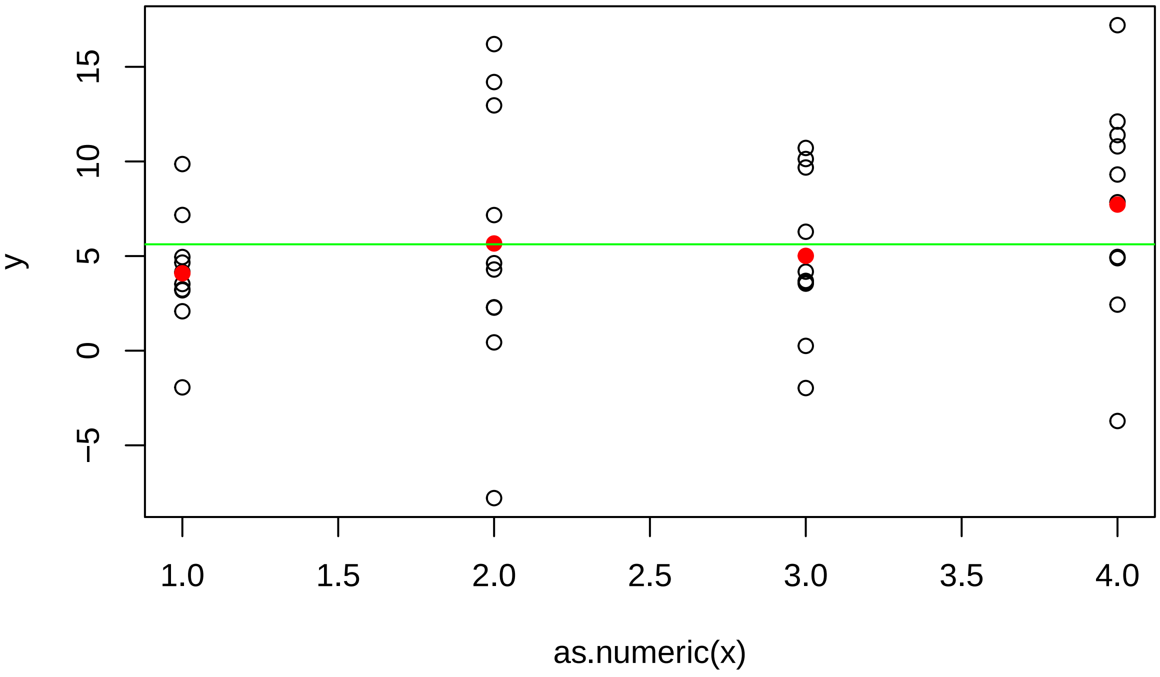
σ^2 is also what MSE is approximating (always, as we’ve seen).

Let’s make sense of this via simulation.

```r
mus <- c(5,5,5,5)  #the null is true
mstr <- c()
pvals <- c()
for (i in 1:100) {
  y <- rep(mus, 10) + rnorm(40, 0, 5) #sigma^2=25
  x <- as.factor(rep(c('a','b', 'c', 'd'), 10))
  fit <- aov(y~x)
  mstr[i] <- anova(fit)$Mean[1]
  pvals[i] <- anova(fit)$Pr[1]
}
```

Here’s the last data set out of 100 we just generated under the null

```r
data.sim <- data.frame(x,y)
plot(as.numeric(x), y)
means <- mean(~y|x, data=data.sim)
points(1:4, means, pch=19, col='red')
abline(mean(y), 0, col='green')
```

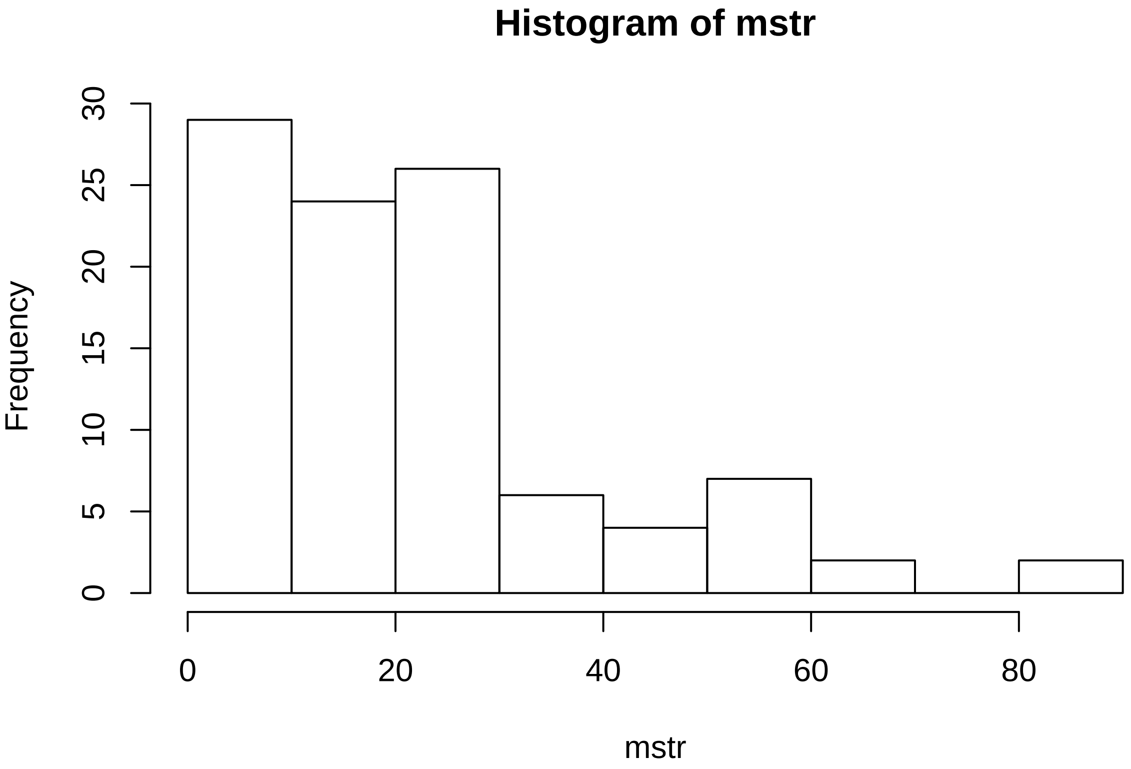


```
mstr[100]
```

```
## [1] 23.77191
```

The MSTR values can fluctuate pretty dramatically, depending on the data you get. Let's see how varied they are, and where they are centered.

```
hist(mstr)
```



```
mean(mstr)
```

```
## [1] 22.23922
```

Right about where we'd expect it to be centered. Sometimes, it's seemingly not even close. But this behavior is well understood, and we account for that with the F test (which is pretty darn similar to all the F tests we've seen already)

$$F = MSTR/MSE \qquad \text{where} \qquad df(MSE) = \sum_i n_i - k$$

To be fair, the data we generated originally (where the treatments did differ, but the data was noisy), has an MSTR of

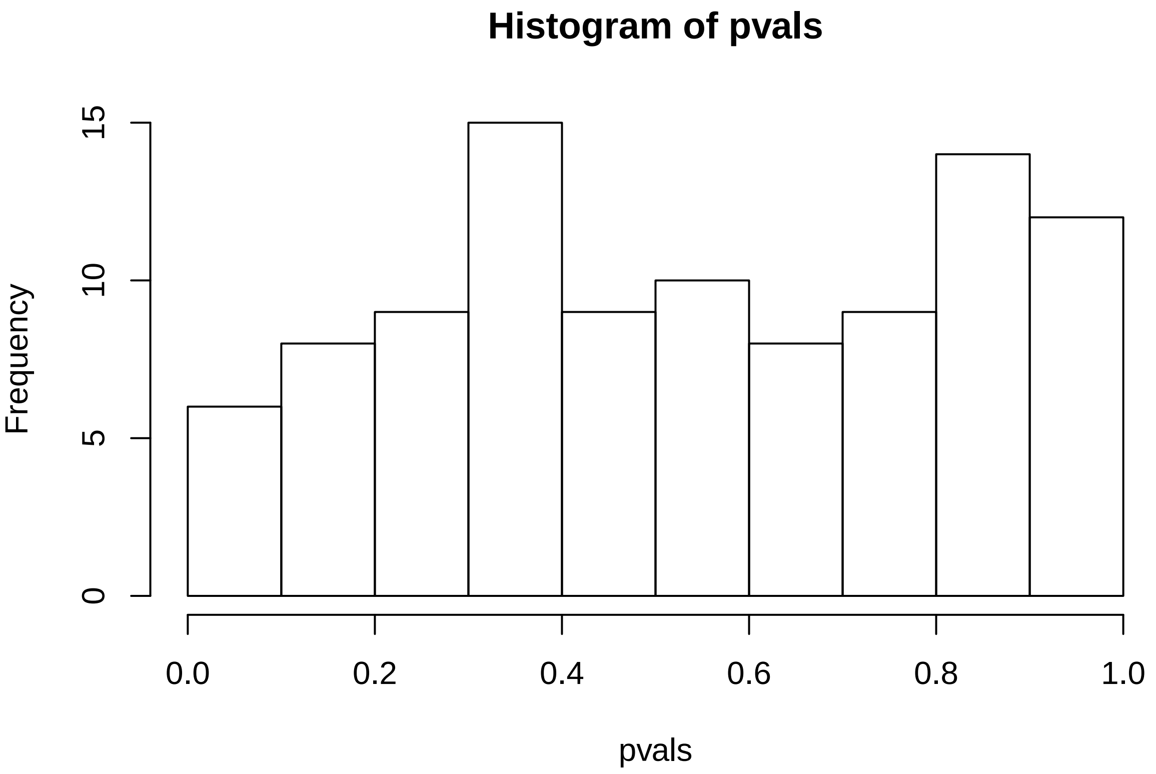
```
anova(aov(y1~x))$Mean[1]
```

```
## [1] 15.55492
```

The original (nice) data has a different value of σ^2, so it can't be compared with this simulation.

Here's justification that the F test expects the vast majority (95%) of these fluctations.

```
hist(pvals)
```



```
mean(pvals < .05)
```

```
## [1] 0.03
```

For you to do:

We use aov() (instead of lm, but otherwise things look similar) to fit the model, and then to get the ANOVA table, which includes the sum/mean squares, F statistic and p-value, you run anova() on that aov object.

The data set iris (data(iris)) contains 4 different measurements of an iris plant over 3 different species. Answer the following and justify with R output (expect e).

a) Is the design balanced? Looks balanced (50 each in group setosa, versicolor, virginica).

b) What are the n_i? 50

c) What is k? 3

d) If you were only allowed a single one of these numeric variables, and you had to guess the species based on that variable, which of those 4 variables would you choose? Justify via F statistics.

```
data(iris)

fit1<-aov(Sepal.Length~Species, data=iris)
anova(fit1)$Pr[1]
```

```
## [1] 1.669669e-31
```

```
fit2<-aov(Sepal.Width~Species, data=iris)
anova(fit2)$Pr[1]
```

```
## [1] 4.492017e-17
```

```
fit3<-aov(Petal.Length~Species, data=iris)
anova(fit3)$Pr[1]
```

```
## [1] 2.856777e-91
```

```
fit4<-aov(Petal.Width~Species, data=iris)
anova(fit4)$Pr[1]
```

```
## [1] 4.169446e-85
```

Petal.Length is the most significant based on F statistics

e) Are you sure this is the right way to answer this question?

Choosing the variable by largest F statistic ensures say that Versicolor is different from Virginica and Setosa, but does not ensure that all the variables are different (so you can discriminate between them well).