The analysis of variance

Guillaume Filion

September 26, 2011

1 The problem

In 1992, Denke and Grundy set out to study the effect of different insaturated fats in the diet on the total blood cholesterol. They controlled the diet of 15 patients with the same amount of insaturated fats of different kinds. Here is what they obtained (the study is real but the data is fake).

Total blood cholesterol (mmol/L)		
High oleic	High lauric	High palmitic
3.08	3.75	2.83
5.71	5.13	3.91
5.28	5.74	6.46
5.26	3.22	7.34
4.44	5.51	6.36

In your R session, manually enter the data in 3 vectors of length 5 (`oleic`, `lauric`, `palmitic`).

```
> oleic <- c(3.08, 5.71, 5.28, 5.26, 4.44);
> lauric <- c(3.75, 5.13, 5.74, 3.22, 5.51);
> palmitic <- c(2.83, 3.91, 6.46, 7.34, 6.36);
```

Exercise 1

Can we assume that the data is Gaussian? What is the null hypothesis in that case?

Exercise 2

What are the risks of doing 3 t tests? Find a statistic F that allows you to do a single test. Formulate the null hypothesis.

2 The test statistic

As usual, we need to make our life easier by writing a function to compute the statistic and then resample the statistic under the null hypothesis.

Exercise 3

Write a function that computes the F statistic. Assume that the input is a `list` of vectors. Try to write a function that would work for any number of samples, not just 3.

Exercise 4

Using the function to compute the F statistic or from the formula, prove that F, like the effect size t is invariant by translation and scaling. What does that mean for the resampling?

Exercise 5

Resample F under the null hypothesis. Find estimates of the limits of the rejection region. Finish the test. Estimate the p-value.

Exercise 6

Compare your results with that of `R`. The way to do an ANOVA in `R` is not very intuitive. You need to put all the observations in a single vector, say `obs`, build a `factor` that indicates which sample they belong to, say `smpl` and use `anova(lm(obs ~ smpl))`.

Answer of Exercise 1

Yes, we can assume that the data is Gaussian. Medical 'reaction norms' often have a Gaussian distribution. The null hypothesis resembles that of the t test. The only difference is that we have more than one population.

1. All datasets are sampled from a Gaussian distribution.
2. The parameters are unknown, but equal in all cases.
3. Sampling is IID.

Answer of Exercise 2

The risks of doing several t tests are that the power will be low (because we will not use all the samples in each test) and that we will need to correct for multiple testing (and the tests are not independent by the way).

A good test statistic is the ratio of the variance of the means to the mean of the variances. For simplicity, we drop the normalization constants and work with the sum of squares instead.

$$F = \frac{\sum_{i=1}^{3}\sum_{j=1}^{5}(\bar{x}_i - \bar{x})^2}{\sum_{i=1}^{3}\sum_{j=1}^{5}(x_{ij} - \bar{x}_i)^2} \tag{1}$$

It is associated with the alternative hypothesis in which the item 2 is replaced by 'At least two samples are taken from distributions with a different mean'.

Answer of Exercise 3

```
> F <- function(samplist) {
+   grandmean <- mean(unlist(samplist));
+   num <- 0;
+   denom <- 0;
+   for (smpl in samplist) {
+     # Terms in j are ientical in numerator of formula (1).
+     num = num + length(smpl)*(mean(smpl)-grandmean)^2;
+     # Compute sum of squares as (n-1)*var.
+     denom = denom + (length(smpl)-1)*var(smpl);
+   }
+   return (num/denom);
+ }
> F.obs <- F(list(oleic, lauric, palmitic));
> F.obs;

[1] 0.06273586
```

Answer of Exercise 4

The demonstration is essentially the same as the one shown for the t statistic. We can verify it with a few examples.

```
> F(list(oleic+1, lauric+1, palmitic+1));

[1] 0.06273586
```

```
> F(list(pi*oleic+1, pi*lauric+1, pi*palmitic+1));
```

```
[1] 0.06273586
```

This means that we can resample F with standard Gaussian variables, because all Gaussian variables will have the same distribution of F under the null hypothesis.

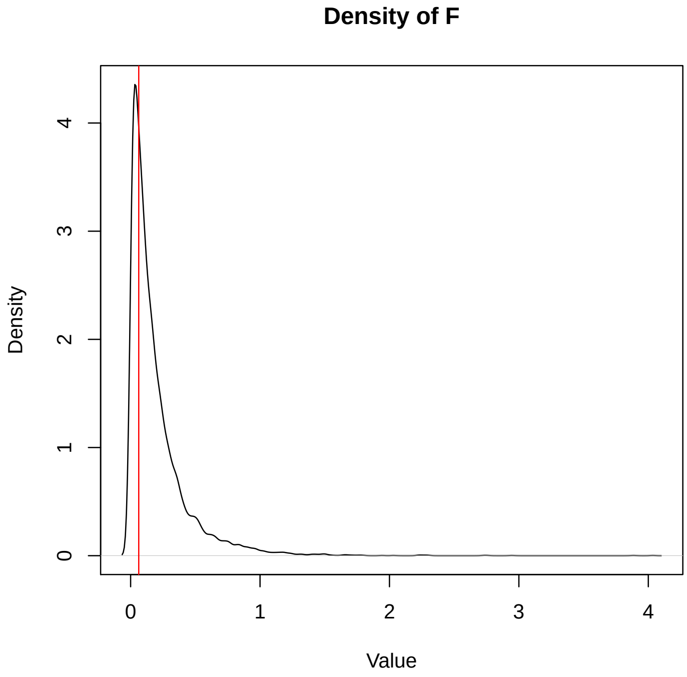
Answer of Exercise 5

```
> F.smpl <- rep(NA, 10000);
> for (i in 1:10000) {
+   F.smpl[i] <- F(list(rnorm(5), rnorm(5), rnorm(5)));
+ }
> plot(density(F.smpl), main="Density of F", xlab="Value");
> # Note that the rejection region is necessarily unilateral.
> abline(v=F.obs, col=2);
> quantile(F.smpl, probs=0.95);
```

```
      95%
0.6540971
```

```
> # Not significant. Let's have a look at the p-value.
> mean(F.smpl >= F.obs);
```

```
[1] 0.6974
```



Note that the distribution of F is not symmetric. If F is small, all the samples have similar means and we do not want to reject the null hypothesis.

We reject it only when F is large, so in the analysis of variance, the test is only one-sided.

Answer of Exercise 6

```
> obs <- c(oleic, lauric, palmitic);
> smpl <- factor(rep(c("oleic", "lauric", "palmitic"), each=5));
> anova(lm(obs ~ smpl));

Analysis of Variance Table

Response: obs
          Df  Sum Sq Mean Sq F value Pr(>F)
smpl       2  1.5051 0.75253  0.3764 0.6941
Residuals 12 23.9903 1.99919
```

Again, the F used by `R` is not the same as the one we have used (it varies only by a scaling factor), but the p-value, indicated in the column `Pr(>F)` is close to the one we found.