

Contents

1	Comparing multiple populations	1
1.1	Sample data from k treatment groups	1
1.2	Treatment means and grand mean	2
1.3	Sums of squares	2
1.4	f distribution	3
1.5	Hypothesis test	3
1.6	One-way ANOVA summary table	3
1.7	Complete R code for one-way ANOVA	3
2	A few formulas for simple calculations	6

1 Comparing multiple populations

We now wish to compare k populations. We have a sample from each population and wish to determine if these populations are actually different. From the i^{th} population we have a sample of size n_i with $n = \sum_{i=1}^k n_i$ the total size of the entire dataset. It is best if all the sample sizes are the same or at least of the same order of magnitude.

We will refer to each population as a “treatment” as is customary in the analysis of variance world. This is because this technique is often used to compared different treatments, e.g. fertilizers or medical procedures.

1.1 Sample data from k treatment groups

A dataset from multiple populations/groups/treatments:

$$\begin{array}{ll}
 \text{treatment 1 sample} & X_{11}, X_{12}, \dots, X_{1n_1} \\
 \text{treatment 1 sample} & X_{21}, X_{22}, \dots, X_{2n_2} \\
 & \vdots \\
 \text{treatment } k \text{ sample} & X_{k1}, X_{k2}, \dots, X_{kn_k}
 \end{array}$$

We wish to know if these samples are actually taken from the same population or not. The assumption is

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where $\mu_i = \mu + \alpha_i$ is the mean for treatment i and ϵ_{ij} is the random deviation from this mean for data point X_{ij} . We are primarily interested in testing whether or not $\alpha_i = 0$ for all treatments. We will test the null hypothesis that $\alpha_i = 0$ for all i under the assumption that $\epsilon_{ij} \sim N(0, \sigma^2)$.

1.2 Treatment means and grand mean

Here are the summary statistics:

A dataset from multiple populations/groups/treatments:

$$\begin{aligned}
 \text{treatment mean 1} \quad \bar{x}_{1.} &= \sum_{j=1}^{n_1} X_{1j} \\
 \text{treatment mean 2} \quad \bar{x}_{2.} &= \sum_{j=1}^{n_2} X_{2j} \\
 &\vdots \\
 \text{treatment mean } k \quad \bar{x}_{k.} &= \sum_{j=1}^{n_k} X_{kj} \\
 \text{grand mean} \quad \bar{x}_{..} &= \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij}
 \end{aligned}$$

1.3 Sums of squares

We will now define a few different sums of squares:

$$\begin{aligned}
 \text{sum of squares total} \quad SST &= \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2 \\
 \text{sum of squares treatments} \quad SS(Tr) &= \sum_{i=1}^k n_i (\bar{x}_{i.} - \bar{x}_{..})^2 \\
 \text{sum of squared errors/residuals} \quad SSE &= \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2 \\
 SST &= SS(Tr) + SSE
 \end{aligned}$$

Each sum of squares has a different degrees of freedom associated with it. There are n total data points so SST has $n - 1$ degrees of freedom, there are k treatments, so $SS(Tr)$ has $k - 1$ degrees of freedom, and SSE has $n - k$ degrees of freedom. Note that the degrees of freedom sum as well: $n - 1 = (k - 1) + (n - k)$.

1.4 f distribution

Under the assumption that $\epsilon_{ij} \sim N(0, \sigma^2)$ and are independent, then we can get two different estimators for σ^2 :

$$\hat{\sigma}^2 = \frac{SS(Tr)}{k-1}$$
$$\hat{\sigma}^2 = \frac{SSE}{n-k}$$

It follows that

$$\left(\frac{SS(Tr)}{(k-1)\sigma^2} \right) / \left(\frac{SSE}{(n-k)\sigma^2} \right)$$

is f -distributed with $k-1$ numerator degrees of freedom and $n-k$ denominator degrees of freedom.

We define the mean squared errors:

$$MS(Tr) = \frac{SS(Tr)}{k-1}$$
$$MSE = \frac{SSE}{n-k}$$

and thus $\frac{MS(Tr)}{MSE} \sim f_{k-1, n-k}$.

We use this to create a test statistics and conduct a hypothesis test.

1.5 Hypothesis test

$H_0 : \mu_i = \mu_j$ for all i, j (all treatments have the same mean)

$H_a : \mu_i \neq \mu_j$ for at least one i, j pair (some treatment differs from another)

Recalling that $X_{ij} = \mu + \alpha_i + \epsilon_{ij}$ the hypotheses can also be rephrased as

$H_0 : \alpha_i = 0$ for all i

$H_a : \alpha_i \neq 0$ for at least one i

Our test statistic is

$$f^* = \frac{MS(Tr)}{MSE}$$

which has f -distribution with $k-1$ numerator degrees of freedom and $n-k$ denominator degrees of freedom.

The p -value is $P(f_{k-1, n-k} \geq f^*) = 1 - \text{pf}(f^*, k-1, n-k)$.

1.6 One-way ANOVA summary table

1.7 Complete R code for one-way ANOVA

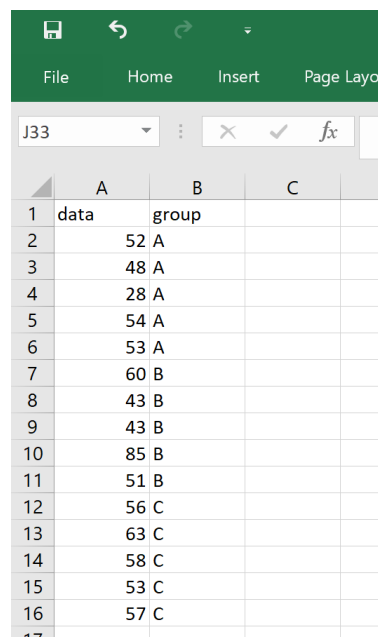
The following is a complete code for conducting a one-way analysis of variance.

The simplest method is to record your data in spreadsheet with two columns, one with the numerical data and the other with the group names. It is important to have actual group names instead of numerical values, e.g. {group1, group2, group3, group4, group5} or {A, B, C, D, E} instead of {1, 2, 3, 4, 5}. See image below. You can change the data header from “data” to whatever you like (no spaces) and the group header from “group” to whatever you like. Of course, you can change the group names from {A, B, C} to whatever you like as well.

Here is dataset with 3 treatments and 5 sampled points from each treatment:

#	data	group
1	52	A
2	48	A
3	28	A
4	54	A
5	53	A
6	60	B
7	43	B
8	43	B
9	85	B
10	51	B
11	56	C
12	63	C
13	58	C
14	53	C
15	57	C

Here it is in an Excel spreadsheet:



For this spreadsheet, I would highlight the cells A1:B16, copy them to the computer’s clipboard, and then execute the commands in R:

```
> d = read.data("clipboard",header=TRUE)
summary(aov(d$data~d$group))
```

It gives output

```
          Df Sum Sq Mean Sq F value Pr(>F)
d$group    2  329.2   164.6   1.132  0.354
Residuals 12 1744.4   145.4
```

The test statistics is $f^* = 1.132$ and the p -value is $p = 0.354$ thus we would not reject the null hypothesis for this particular dataset. We do not have evidence against the claim that treatments A, B, and C come from the same population.

Here is a complete R code that does the entire analysis of variance hypothesis “manually” so-to-speak. This code requires that you have your data copied to clipboard with headers.

```
> d=read.table("clipboard",header=TRUE)
xdd=mean(d$data)
SST=sum((d$data-xdd)^2)
g=levels(d$group)
k=length(g)
n=length(d$data)
nvec=as.numeric(table(d$group))
SSTr=0
xd=vector("numeric",length=length(g))
for (i in 1:length(g)){
  xd[i]=mean(d$data[d$group==g[i]])
  SSTr=SSTr+nvec[i]*(xd[i]-xdd)^2
}
xdvec=NULL
for (i in 1:length(g)){
  xdvec=c(xdvec,rep(xd[i],nvec[i]))
}
SSE=sum((d$data-xdvec)^2)
MSTr=SSTr/(k-1)
MSE=SSE/(n-k)
f=MSTr/MSE
1-pf(f,k-1,n-k)
```

Here is a complete R code that does the entire analysis of variance hypothesis where you enter each treatment sample data as a separate list. The parts of the code that you must modify are given in red.

```
> x1=c(input group 1 data)
x2=c(input group 2 data)
:
xk=c(input group k data)
data=c(x1,x2,...,xk)
```

```

group=c(rep("A1",length(x1)),rep("A2",length(x2)),...,rep("Ak",length(xk)))
group=as.factor(group)
xdd=mean(data)
SST=sum((data-xdd)^2)
g=levels(group)
k=length(g)
n=length(data)
nvec=as.numeric(table(group))
SSTr=0
xd=vector("numeric",length=length(g))
for (i in 1:length(g)){
  xd[i]=mean(data[group==g[i]])
  SSTr=SSTr+nvec[i]*(xd[i]-xdd)^2
}
xdvec=NULL
for (i in 1:length(g)){
  xdvec=c(xdvec,rep(xd[i],nvec[i]))
}
SSE=sum((data-xdvec)^2)
MSTr=SSTr/(k-1)
MSE=SSE/(n-k)
f=MSTr/MSE
1-pf(f,k-1,n-k)

```

2 A few formulas for simple calculations

We can calculate the ANOVA quantities simply by hand by summing up the data for each treatment group ($\sum x$), and also summing up all the squared data values for each treatment group ($\sum x^2$).

$$CF = \frac{\left(\sum_{i,j} x_{ij}\right)^2}{n}$$

$$SST = \sum_{i,j} x_{ij}^2 - CF$$

$$SS(Tr) = \sum_{i=1}^k \frac{\left(\sum_{j=1}^{n_i} x_{ij}\right)^2}{n_i} - CF$$

$$SSE = SST - SS(Tr)$$

Then we have degrees of freedom $k - 1$ for $SS(Tr)$ and $n - k$ for SSE , and use this to calculate $MS(Tr)$, MSE , and f^* as already discussed.

Example: Here is a small dataset from 3 treatment groups with summary information.

group	A	B	C
data	3	2	3
	4	5	8
	5	5	
	7		
$\sum_{j=1}^{n_i} x_{ij}$	12	19	11
$\sum_{j=1}^{n_i} x_{ij}^2$	50	103	73
n_i	3	4	2

$$CF = \frac{12^2+19^2+11^2}{3+4+2} = \frac{1764}{9} = 196$$

$$SST = 50 + 103 + 73 - 196 = 30$$

$$SS(Tr) = \frac{12^2}{3} + \frac{19^2}{4} + \frac{11^2}{2} - 196 = 2.75$$

$$SSE = 30 - 2.75 = 27.25$$

$$MS(Tr) = \frac{2.75}{2} = 1.375$$

$$MSE = \frac{27.25}{6} \approx 4.54167$$

$$f^* = \frac{1.375}{4.54167} \approx 0.30275$$

$p = 1 - \text{pf}(0.30275, 2, 6) \approx 0.7494396$. Thus we would not reject the null hypothesis in this case.

To work this example using the R code provided, we would type:

```
> x1=c(3,4,5)
  x2=c(2,5,5,7)
  x3=c(3,8)
  data=c(x1,x2,x3)
  group=c(rep("A1",length(x1)),rep("A2",length(x2)),rep("A3",length(x3)))
  group=as.factor(group)
  xdd=mean(data)
  SST=sum((data-xdd)^2)
  g=levels(group)
  k=length(g)
  n=length(data)
  nvec=as.numeric(table(group))
  SStr=0
  xd=vector("numeric",length=length(g))
  for (i in 1:length(g)){
    xd[i]=mean(data[group==g[i]])
    SStr=SSTr+nvec[i]*(xd[i]-xdd)^2
  }
  xdvec=NULL
  for (i in 1:length(g)){
    xdvec=c(xdvec,rep(xd[i],nvec[i]))
  }
```

```
SSE=sum((data-xdvec)^2)
MSTr=SSTr/(k-1)
MSE=SSE/(n-k)
f=MSTr/MSE
1-pf(f,k-1,n-k)
```

which gives the output

```
[1] 0.7494381
```

So both methods agree in p -value to 4 decimal places. If we didn't round our MSE , then we would have matched all decimal places.