Math 321 – Chapter 8 – One-way ANOVA
(draft version 2020/04/29-15:57:40)

Contents

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:

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

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 Assumptions of the test

The null hypothesis assumes that all treatment groups have the same means and variances and are normally distributed. If the actual populations are non-normal or the variances are not equal, then the test will not be as accurate. However, the normality assumption is not as important as long as the data does not have an overabundance of outliers, or has histograms that are roughly clump-shaped. As long as you don't have a good reason to believe the underlying populations are extremely non-normal, then you don't need to worry about that assumption too much.

If the underlying populations have unequal variances, then that can have a more significant impact on the accuracy of the test. Here are a few rough rules-of-thumb to keep in mind on when this test will be reasonably accurate:

1. Sample sizes should be similar, e.g. the largest sample should be no larger than twice the smallest sample.

2. Sample variances should be similar, e.g. the largest variance should be no more than four times the smallest variance.

3. Sample size should scale with variance, e.g. the sample with the largest variance should have the largest sample size.

4. Calculate the quantity

$$f_0 = \frac{n-k}{k-1} \cdot \frac{\sum_{i=1}^{k} s_i^2 - \frac{1}{n}\sum_{i=1}^{k} n_i s_i^2}{\sum_{i=1}^{k} n_i s_i^2 - \sum_{i=1}^{k} s_i^2}$$

   If it is too far away from 1, your sample sizes and variances may be too unbalanced. Ideally this quantity should be between 0.8 and 1.2.

Even these guidelines can be violated and the test still be accurate. It is difficult to give a perfect set of rules that will keep the one-way ANOVA test accurate. Here are some examples to help understand these guidelines. The more one of these guidelines is violated, as long as the others are not, then the test can still be reasonably accurate.

Example 1: $n_1 = 10$, $n_2 = 25$, $n_3 = 15$, $s_1^2 = 1.4$, $s_2^2 = 0.9$, $s_3^2 = 2.3$. The sample sizes don't satisfy the guidelines above, in particular the smallest sample has the largest variance, but $f_0 = 1.125$ and the variances are all well within a factor of 3 of each other. The test should not be too innacurate. In fact, if these given variances are assumed to be the actual population variances for normal populations, then if we desire a 5% level of significance, the actual significance will be 7.1%.

Example 2: $n_1 = 20$, $n_2 = 30$, $n_3 = 40$, $s_1^2 = 0.7$, $s_2^2 = 1.3$, $s_3^2 = 4.1$. The sample sizes satisfy the guidelines above, and scale with the variances, $f_0 = 0.76$, but the largest variance is nearly a factor of 6 times the smallest. The test should not be too innacurate though since most of the guidelines are nearly satisfied. In fact, if these given variances are assumed to be the actual population variances for normal populations, then if we desire a 5% level of significance, the actual significance will be 3.29%.

Example 3: $n_1 = 20$, $n_2 = 30$, $n_3 = 40$, $s_1^2 = 4.1$, $s_2^2 = 1.3$, $s_3^2 = 0.7$. Note that we have taken the sample parameters as the previous example, but have changed the order of the sample sizes. The sample sizes still satisfy the guidelines above, but do not scale with the variances variance, $f_0 = 1.35$, and the largest variance is nearly a factor of 6 times the smallest. The test may be very innacurate since many of the guidelines are not satisfied. In fact, if these given variances are assumed to be the actual population variances for normal populations, then if we desire a 5% level of significance, the actual significance will be 11.6%. If we change the disparity between the variances by setting $s_1^2 = 2.1$, then the accuracy improves vastly to 8% actual significance with $f_0 = 1.2$.

The most important rule of thumb is to have similar variances and then to have similar sample sizes.

1.3 Treatment means and grand mean

Here are the summary statistics:

A dataset from multiple populations/groups/treatments:

$$\text{treatment mean } 1 \qquad \overline{x}_{1\cdot} = \frac{1}{n_1}\sum_{j=1}^{n_1} X_{1j}$$

$$\text{treatment mean } 2 \qquad \overline{x}_{2\cdot} = \frac{1}{n_2}\sum_{j=1}^{n_2} X_{2j}$$

$$\vdots$$

$$\text{treatment mean } k \qquad \overline{x}_{k\cdot} = \frac{1}{n_k}\sum_{j=1}^{n_k} X_{kj}$$

$$\text{grand mean} \qquad \overline{x}_{\cdot\cdot} = \frac{1}{\sum_i n_1}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}$$

1.4 Sums of squares

We will now define a few different sums of squares:

$$\text{sum of squares total} \qquad SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_{..})^2$$

$$\text{sum of squares treatments} \qquad SS(Tr) = \sum_{i=1}^{k} n_i(\overline{x}_{i\cdot} - \overline{x}_{..})^2$$

$$\text{sum of squared errors/residuals} \qquad SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_{i\cdot})^2$$

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

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: $n-1 = (k-1) + (n-k)$.

1.5 f distribution

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

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

$$\hat{\sigma}^2 = \frac{SSE}{n-k}$$

It follows that

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

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.

1.6 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-pf(`f^*`,k-1,n-k)`.

1.7 Effect size

The effect size tells us how much variation between the treatment groups is due to their having different means vs just random variation.

$$\eta^2 = \frac{SS(Tr)}{SST}$$

$\eta^2 \leq 0.01$ is considered a small effect,
$\eta^2 \approx 0.06$ is considered a medium effect,
$\eta^2 > 0.14$ is considered a large effect,

You can interpret η^2 as the "percent of variation in the data that is explained by the difference between the treatment means." Similarly you can interpret $\frac{SSE}{SST}$ as the percent of variation in the data that is due to randomness. Note that $\frac{SSTr}{SST} + \frac{SSE}{SST} = 1$.

1.8 One-way ANOVA summary table

1.9 Complete R code for one-way ANOVA

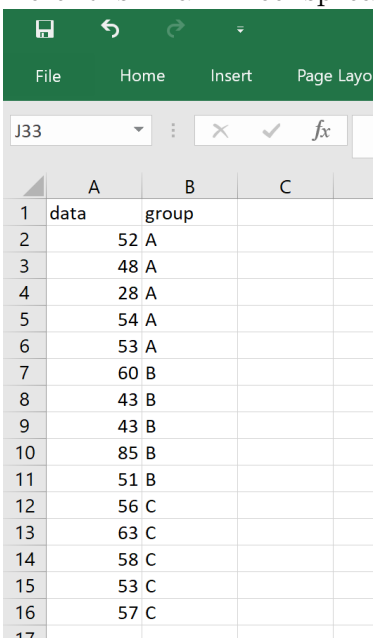
The following is a complete code for conducting a one-way analysis of variance.

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 otuput

```
             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)
  d=d[order(d$group),]
  xdd=mean(d$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 treatmetn 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)
  data=data[order(group)]
  group=group[order(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-pf(0.30275,2,6)` ≈ 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)
  data=data[order(group)]
  group=group[order(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.