ANOVA

MATH 2441, BCIT

Technical Mathematics for Food Technology

May 7, 2018

Analysis of Variance

One-way analysis of variance (ANOVA) is a method of testing the equality of three or more population means by analyzing sample variances. One-way analysis of variance is used with data categorized with one factor (or treatment), so there is one characteristic used to separate the sample data into the different categories.

In order to conduct one-way ANOVA, the requirements are (informally) that

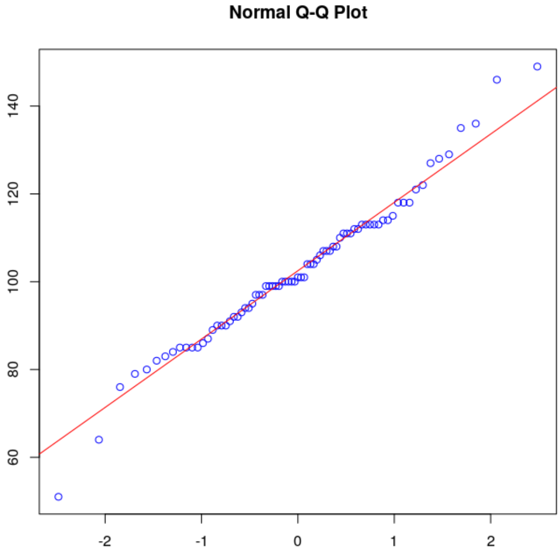
- the samples are independent
- the populations are normally distributed, and
- the population variances are approximately equal.

Analysis of Variance

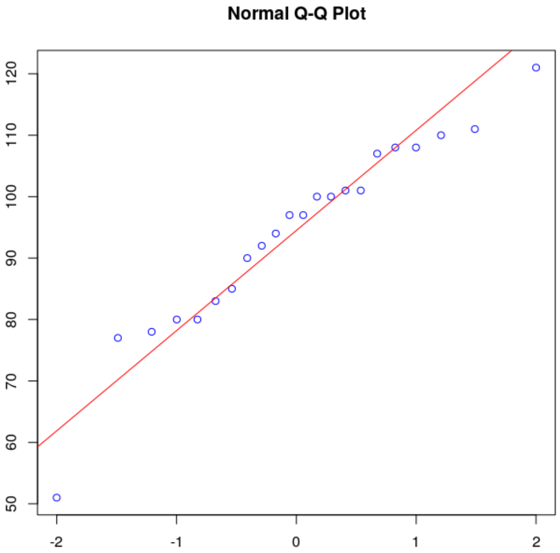
Consider the following R Statistics code. It corresponds to the narrative in Triola, page 562. The result is a test statistic following the *F*-distribution and a *p*-value that can be compared to the significance level. This type of ANOVA is always a right-hand one-tailed test.

```
l<-c(85,90,107,85,100,97,101,64,111,100,76,136,100,90,135,104,149,99,107,
  99,113,104,101,111,118,99,122,87,118,113,128,121,111,104,51,100,113,
  82,146,107,83,108,93,114,113,94,106,92,79,129,114,99,110,90,85,94,127,
  101,99,113,80,115,85,112,112,92,97,97,91,105,84,95,108,118,86,89,100)
m<-c(78,97,107,80,90,83,101,121,108,100,110,111,97,51,94,80,101,92,100,
  77,108,85)
h<-c(93,100,97,79,97,71,111,99,85,99,97,111,104,93,90,107,108,78,95,78,
  86)
n<-c(length(l),length(m),length(h))
group<-rep(1:3,n)
y<-c(l,m,h)
data<-data.frame(y=y,group=factor(group))
fit<-lm(y˜group,data)
anova(fit)
```

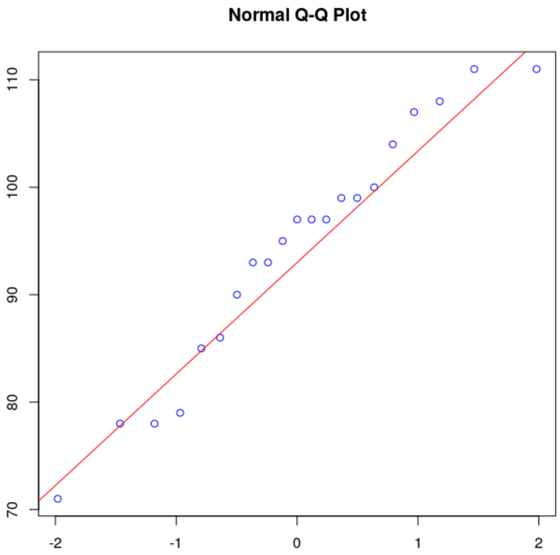
Anova: Lead and Intelligence Quotients (low)



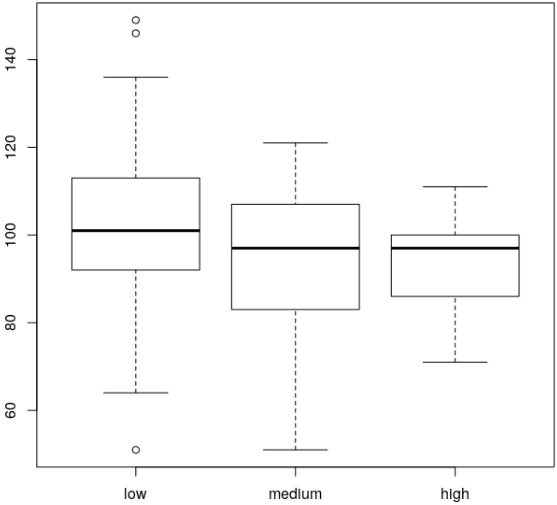
Anova: Lead and Intelligence Quotients (medium)



Anova: Lead and Intelligence Quotients (high)



Anova: Lead and Intelligence Quotients (boxplots)



Anova: Lead and Intelligence Quotients

R Statistics yields the following output:

```
Analysis of Variance Table

Response: y
           Df  Sum Sq Mean Sq F value Pr(>F)
group       2  1920.9  960.45  3.8646 0.0237 *
Residuals 117 29077.1  248.52
---
Signif. codes:  0 ***’ 0.001 **’ 0.01 *’ 0.05 .’ 0.1  ’ 1
```

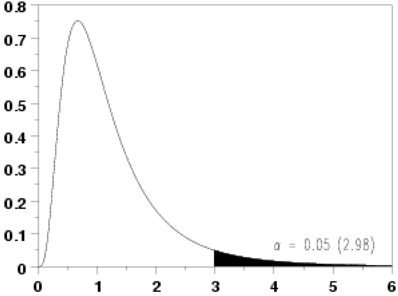
What is important to us is the test statistic, whose distribution is the *F*-distribution, 3.8646; and the *p*-value 0.0237. Since ANOVA is always one-tailed, area to the right, all you need to do is compare the *p*-value to the significance level. “If *p* is low, the NULL must go.”

F-Distribution

The F-distribution depends on two different degrees of freedom, which makes using a table of critical values awkward. We will use p-values provided by technology instead. A table with critical values is here:

`http://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm`.

The shape of the F-distribution is similar to the shape of the χ^2-distribution.



Degrees of Freedom for the F-Distribution

A variable that follows an F-distribution is the ratio of two independent chi-square variables divided by their respective degrees of freedom. Therefore, the F-distribution has two different degrees of freedom, one for the numerator and one for the denominator.

- Are all of the data values within any one group the same? No! So there is some within group variation. The within group is sometimes called the error group.
- Are all the sample means between the groups the same? No! So there is some between group variation. The between group is sometimes called the treatment group.

Degrees of Freedom for the F-Distribution

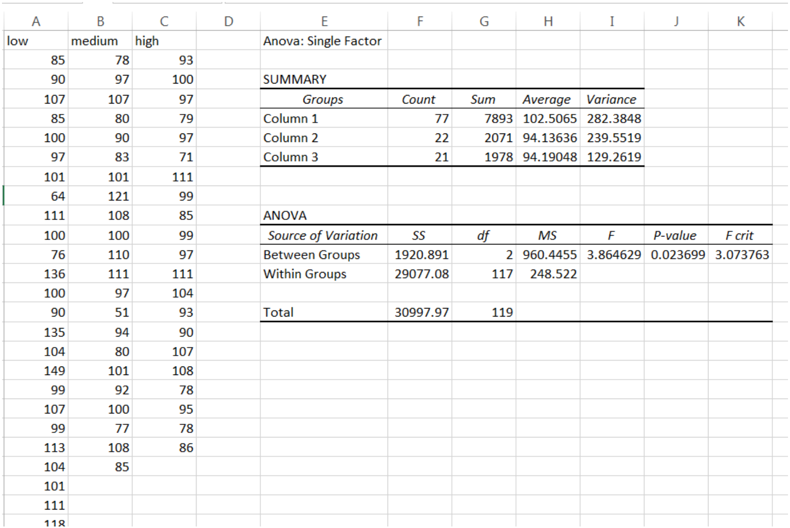
There are two sources of variation here. The between group and the within group. Let there be k groups and N data points within those k groups.

- The between group degree of freedom (for the numerator) is $k - 1$, just as it was for a categorical goodness-of-fit test.
- The within group degree of freedom (for the denominator) is $N - k$, just as it was for linear regression with $k = 2$.

In the lead example, there are three groups (low, medium, high) with 77, 22, and 21 individual data points respectively. The first degree of freedom is $3 - 1 = 2$; the second degree of freedom is $120 - 3 = 117$. A 0.05 significance test would result in a critical value of `qf(0.95,2,117)` which equals 3.073763 (this critical value is also displayed in Excel). We know from the ANOVA performed in Excel and in R Statistics that the test statistic is 3.8646. We reject the null hypothesis.

One-Way ANOVA in Microsoft Excel

For instructions, see `http://www.excel-easy.com/examples/anova.html`.



A	B	C	D	E	F	G	H	I	J	K
low	medium	high		Anova: Single Factor						
85	78	93								
90	97	100		SUMMARY						
107	107	97		*Groups*	*Count*	*Sum*	*Average*	*Variance*		
85	80	79		Column 1	77	7893	102.5065	282.3848		
100	90	97		Column 2	22	2071	94.13636	239.5519		
97	83	71		Column 3	21	1978	94.19048	129.2619		
101	101	111								
64	121	99								
111	108	85		ANOVA						
100	100	99		*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
76	110	97		Between Groups	1920.891	2	960.4455	3.864629	0.023699	3.073763
136	111	111		Within Groups	29077.08	117	248.522			
100	97	104								
90	51	93		Total	30997.97	119				
135	94	90								
104	80	107								
149	101	108								
99	92	78								
107	100	95								
99	77	78								
113	108	86								
104	85									
101										
111										
118										

ANOVA Exercise

Exercise 1: Copy and past the comma-separated value data (in cm) on the next slide. You have eight graduate students. They each measure the height of kindergarten children (six years of age, when there is no significant height difference between girls and boys). Can you rely on their measuring techniques to be consistent?

ANOVA Exercise

```
a<-c(119.8,116.5,120.3,111.2,107.6,116.1,110.0,114.9,118.7,121.7,115.6,109.4,
107.8,113.9,114.7,118.9,124.4,109.1,111.9,117.6,111.1,121.8,116.7,117.9,117.9)
b<-c(114.0,112.4,117.0,117.0,116.2,121.1,118.6,119.7,118.4,122.5,122.0,117.8,
124.2,112.8,116.4,112.2,113.7,114.7,116.8,122.4,111.7,115.9,114.4,109.6,123.2,
110.5,109.6,126.4)
c<-c(112.0,109.3,125.1,101.9,116.0,113.6,117.4,122.6,114.0,118.0,111.3,118.4,
117.0,121.3,118.6,119.8,120.0,121.5,122.3,120.7,123.2,106.3,121.8,123.2)
d<-c(116.3,110.5,109.1,117.5,124.6,108.1,117.9,111.7,126.9,110.9,115.7,110.8,
112.1,114.7,123.1,119.3,109.4,112.3,112.3,120.2,119.3,104.7,113.6,112.4,115.2,
112.9,121.8,128.8,115.7,114.6)
e<-c(123.1,119.2,113.3,114.2,110.3,120.6,111.4,119.7,106.7,112.1,113.3,119.3,
119.5,125.6,120.3,110.9,112.3,118.1,115.7,112.6,113.3,115.4,121.8,116.7,109.1,
117.8,113.0,107.4,117.3)
f<-c(114.1,116.5,114.5,108.4,111.6,118.1,116.6,114.0,116.3,109.6,117.0,112.0,
116.7,121.6,119.8,114.3,118.5,121.4,110.9,110.6,118.5,109.8,123.5,119.9,114.1,
115.2,123.8,113.6,110.3,111.7,106.6)
g<-c(111.9,116.4,111.8,105.8,111.9,117.1,113.8,111.9,107.5,111.7,116.3,118.4,
116.0,117.3,126.3,114.8,113.1,120.6,105.6,119.6,113.6,116.9,120.7,121.5,117.1,
117.4)
h<-c(114.4,120.0,115.1,121.6,123.5,120.3,107.8,110.7,112.6,115.6,110.2,116.3,
121.1,112.4,123.6,116.1,116.2,117.3,116.3,113.1,116.0,111.4,111.7,117.1,116.2,
113.9)
```

R Statistics Code

```
n<-c(length(a),length(b),length(c),length(d),
length(e),length(f),length(g),length(h))
group<-rep(1:8,n)
y<-c(a,b,c,d,e,f,g,h)
data<-data.frame(y=y,group=factor(group))
fit<-lm(y˜group,data)
anova(fit)
```

The result is

```
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value Pr(>F)
group       7  117.2  16.747  0.7035 0.6691
Residuals 211 5023.4  23.807
```

ANOVA Exercise

Exercise 2: Susan predicts that students will learn most effectively with a constant background sound, as opposed to an unpredictable sound or no sound at all. She randomly divides twenty-four students into three groups of eight. All students study a passage of text for 30 minutes. Those in group 1 study with background sound at a constant volume in the background. Those in group 2 study with noise that changes volume periodically. Those in group 3 study with no sound at all. After studying, all students take a 10 point multiple choice test over the material. Test the appropriate null hypothesis using one-way ANOVA at a 0.05 significance level.

constant sound	7	4	6	8	6	6	2	9
random sound	5	5	3	4	4	7	2	2
no sound	2	4	7	1	2	1	5	5

Final Exam Practice

Exercise 3: At a gas station, 40% of customers fill their tanks. Of those who fill their tanks, 80% pay with a credit card.

1. What is the probability that a customer fills their tank and pays with a credit card?
2. What is the probability that either three or four out of ten customers fill their tank and pay by credit card?
3. What is the probability that more than half of eight customers fill their tank and pay by credit card?

Final Exam Practice

Exercise 4: At a certain time in the afternoon, London Heathrow sees on average 2 planes landing per minute.

1. What is the probability of four or more planes landing in one minute?
2. What is the probability that no plane will land in a particular minute?

Final Exam Practice

Exercise 5: A classic story involves four carpooling students who missed a test and gave as an excuse a flat tire. On the makeup test, the instructor asked the students to identify the particular tire that went flat. If they really didn't have a flat tire, would they be able to identify the same tire? The author asked 41 other students to identify the tire they would select. The results are listed in the following table (except for one student who selected the spare). Use a 0.05 significance level to test the author's claim that the results fit a uniform distribution. What does the result suggest about the ability of the four students to select the same tire when they really didn't have a flat?

left front	11
right front	15
left rear	8
right rear	6

Final Exam Practice

Exercise 6: The police department in Madison, Connecticut, released the following numbers of calls for the different days of the week during a recent February that had 28 days: Monday (114); Tuesday (152); Wednesday (160); Thursday (164); Friday (179); Saturday (196); Sunday (130). Use a 0.01 significance level to test the claim that the different days of the week have the same frequencies of police calls. Is there anything notable about the observed frequencies?

Final Exam Practice

Exercise 7: Listed below are numbers of enrolled students (in thousands) and numbers of burglaries for randomly selected large colleges in a recent year (based on data from the New York Times). Find the best predicted number of burglaries for Ohio State, which had an enrollment of 51,800 students. Does a 95% confidence interval for the predicted value contain 329, which was the actual number of burglaries?

```
enrolment<-c(32,31,53,28,27,36,42,30,34,46)
burglaries<-c(103,103,86,57,32,131,157,20,27,161)
```

Final Exam Practice

Exercise 8: Researchers measured skulls from different time periods in an attempt to determine whether interbreeding of cultures occurred. Results are given below (based on data from "Ancient Races of the Thebaid," by Thomson and Randall-Maciver, Oxford University Press). Use a 0.01 significance level to test the claim that the mean maximal skull breadth in 4000 BCE is less than the mean in 150 CE.

4000 BCE (Maximal Skull Breadth)	$n = 30$	$\bar{x} = 131.37mm$	$s = 5$
150 CE (Maximal Skull Breadth)	$n = 30$	$\bar{x} = 136.17mm$	$s = 5$

Final Exam Practice

Exercise 9: Listed below are speeds (mi/h) measured from southbound traffic on 1-280 near Cupertino, California (based on data from SigAlert). This simple random sample was obtained at 3:30 p.m. on a weekday. Use a 0.05 significance level to test the claim that the sample is from a population with a mean that is less than the speed limit of 65 mi/h.

```
speed<-c(62,61,61,57,61,54,59,58,59,69,60,67)
```

Final Exam Practice

Exercise 10: In a presidential election, 308 out of 611 voters surveyed said that they voted for the candidate who won. Use a 0.10 significance level to test the claim that along all voters, the percentage who believe that they voted for the winning candidate is equal to 43%, which is the actual percentage of votes for the winning candidate.

End of Lesson

Next Lesson: End of Term! Have a Happy Holiday!