

ANOVA: analysis of variance

Lada Adamic
March 11, 2008
SI 544

1 one-way ANOVA

First we will work with Prof. Karen Markey's data on people's understanding of library subject headings (for complete info see <http://www.si.umich.edu/~ylime/NewFiles/morekmd.html#Anchor-Subject-54980>). I'll be passing out some of the surveys for you to look at. The surveys were conducted at 3 Michigan public libraries. Each person taking the survey was asked to fill out demographic info such as age, sex, occupation, etc., as well as to try and write a description for 8 library subject headings: e.g. "Cattle ~ United States ~ Marketing". All subjects at each library were asked to interpret the same 8 headings, for a total of 24 headings across all three libraries. There were three ways the subject headings were presented: just the heading itself, the heading along with the headings directly preceding and following them in alphabetical order, and the headings in a particular book (bibliographic) record. Sometimes the headings were kept in their original order, and sometimes they were rearranged to a recommended standardized order.

```
# load the demographic information for each person taking the survey
demog = read.table("oclc demographics.txt", head=T, sep='\t')

# load the results of the survey
surveyresults = read.table("dataoclcpub.txt", head=T, sep='\t', strip.white=T)

# calculate the percent subject headings interpreted correctly
scorebysurvnum = tapply((surveyresults$correct == "c"), surveyresults$survnum, mean)

# combine them into one data frame
demogandscore = data.frame(demog, scorebysurvnum)

# attach it
attach(demogandscore)

# look at what we have
> summary(demogandscore)
  survnum      sex      age      libuse
Min.   : 1.00    f   :205  Min.   : 9.00    a: 19
1st Qu.: 77.75   m   :101  1st Qu.:14.00   b:117
Median :154.50  NA's: 2    Median :18.00   c:119
Mean   :154.50                      Mean   :28.31   d: 41
3rd Qu.:231.25                      3rd Qu.:42.00   e: 12
Max.   :308.00                      Max.   :74.00
                                     NA's   : 5.00

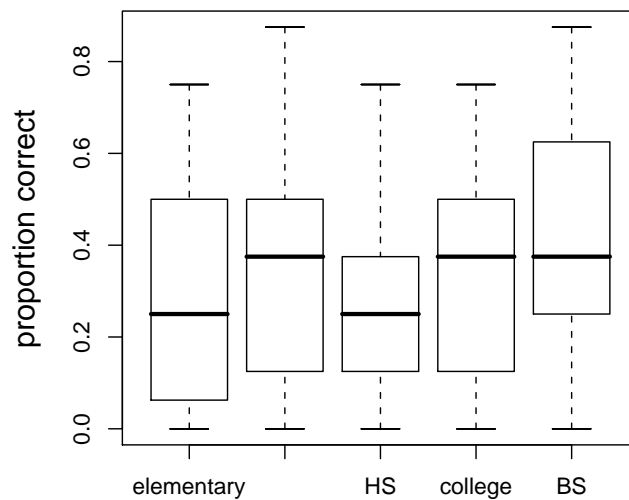
  eductn      profssn  whichlibrary scorebysurvnum
a   :52    student :128    lib1:106      Min.   :0.0000
```

```

b :80  retired : 21  lib2:103  1st Qu.:0.1250
c :30  homemaker: 15  lib3: 99  Median :0.3750
d :55  teacher  : 10  Mean   :0.3449
e :79  clerk    : 6   3rd Qu.:0.5000
NA's:12 (Other)  : 77  Max.   :0.8750
      NA's    : 51  NA's   :1.0000

```

In essence we have 308 people who took the survey, 205 of whom were female, 101 of whom were male. Their ages ranged from 9 to 74, they had varying levels of library use (a:daily, b:weekly, c:monthly, d: 2 to 3 times/yr, e: 1-2 times/yr), with most of them coming to the library on a weekly or monthly basis. There were a lot of students (128), followed by retirees and homemakers, and then a variety of other professions. The score, in terms of the proportion of subject headings which were correctly identified ranges from 0 (a person who got none of the headings correct) to 0.875 (a person who got 7 out of 8 of them correct). No one had gotten all the headings right.



The first thing we'll do is create boxplots for the scores grouped by the education level of the person, ranging from having completed elementary school to having a college degree. From the boxplot, we can see that high school kids (those who have completed JHS, so the JHS box and whisker plot) do about as well as those who have had some college education (the boxplot labeled "college"). Now we'd like to test if any of these means is significantly different from any of the others. Which means that we will be doing an F-test for the null hypothesis that all the means are equal.

```

> anova(lm(scorebysurvnum ~ eductn))
Analysis of Variance Table

Response: scorebysurvnum
      Df Sum Sq Mean Sq F value Pr(>F)
eductn  4  0.5786  0.1447  2.5262 0.04149 *
Residuals 238 13.6290  0.0573
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

We see that we can reject the hypothesis that all the means are equal at the 0.05 level. Great. But which off all the pairs of means are actually different? As we've discussed in class, we can't go and t-test all pairs against each other, because our probability of committing a type-I error (rejecting the null hypothesis when it is true) goes up with each additional test we make. So we need to use a correction, that will multiply the p-value by a factor corresponding to the number of tests made. The Bonferroni adjustment multiplies all p -values, while the Holm method (applied in R by default), corrects the smallest p by the full number of tests, the second smallest by $n - 1$, etc. We will use the `pairwise.t.test()` method for this, which will conveniently do all the pairwise t-tests for us and put them in a nice little table:

1.1 Pairwise t-tests

```
> pairwise.t.test(scorebysurvnum, eductn,p.adj="bonferroni")
```

```
Pairwise comparisons using t tests with pooled SD
```

```
data: scorebysurvnum and eductn
```

	a	b	c	d
b	1.000	-	-	-
c	1.000	1.000	-	-
d	1.000	1.000	1.000	-
e	0.031	0.209	0.068	1.000

```
P value adjustment method: bonferroni
```

```
> pairwise.t.test(demogandscore$scorebysurvnum, demogandscore$eductn)
```

```
Pairwise comparisons using t tests with pooled SD
```

```
data: demogandscore$scorebysurvnum and demogandscore$eductn
```

	a	b	c	d
b	1.000	-	-	-
c	1.000	1.000	-	-
d	1.000	1.000	1.000	-
e	0.031	0.167	0.062	1.000

```
P value adjustment method: holm
```

What we are getting in the table are the p-values for the t-tests between e.g. a and c , multiplied by the number of tests in the case of bonferroni. If this value exceeds 1, then R returns 1.0000 for that t-test. According to the results at the 0.05 level we can only be sure that library partons with a college degree are doing better than kids who have only completed elementary school. For both the Holm and Bonferroni methods (where we penalize different p-scores non-uniformly), we have at the 0.1 level that people with a college degree do better than people with just a high school degree and no years of college.

1.2 Paired vs. pairwise?

One of you asked the very relevant question of what the difference is between a paired and a pairwise t-test. A pairwise t-test means that you have more than 2 groups, and you are pairwise comparing them (meaning that you have $n(n - 1)/2$ comparisons, if you have n groups.

Paired t-tests, if you remember, refer to just two samples, but each observation in one sample is paired with an observation in the other. For example, the two samples could be before and after a treatment was administered to the patient.

Let's roll back to a t-test and see how we can apply it to this data. One of the questions posed in the study was whether the order of the subdivisions within the subject headings mattered.

If the cataloger chooses to apply subdivisions, the subdivisions should always appear in the following order: topical, geographic, chronological, form (Conway 1992, 6). For example, one of the original subject headings was "Education-United States-Finance" and the recommended re-ordering would be "Education-Finance-United States". The ordering changes the likely interpreted meaning.

So let's do the following. For each subject heading, we will take the proportion of people who got it right when it was in the original order, and we will pair it with the proportion of people who got it right when it was in the recommended order, to see if there's a difference.

```
> summary(surveyresults)
  survynum      subjtype sex      libcode      shnum      ordpatrn form      order
Min.   : 1.00  a:1264  -1: 16  Min.   :1.000  Min.   : 1.00  a:1240  a:824  o:1232
1st Qu.: 77.75  c:1200  f :1632  1st Qu.:1.000  1st Qu.: 6.00  b:1224  b:832  r:1232
Median :154.50                m : 816  Median :2.000  Median :12.00                p:808
Mean   :154.50                Mean   :1.977  Mean   :12.32
3rd Qu.:231.25                3rd Qu.:3.000  3rd Qu.:18.00
Max.   :308.00                Max.   :3.000  Max.   :24.00

  question  correct      code1      code2      assess      assest
Min.   :1.00  -1 : 102  ids   :441  ric   : 25  7      :505  (i)ids :433
1st Qu.:2.75  a  : 1   lmo   :350  loi   : 20  4      :461  (i)lmo :336
Median :4.50  c  : 850  loi   :349  rmo   : 16  6      :389  (i)loi :327
Mean   :4.50  i  :1510  ric   :285  lmo   : 11  5      :387  (c)cds :281
3rd Qu.:6.25  NA's: 1   cds   :284  ids   : 9  1      :166  (c)    :261
Max.   :8.00                (Other):493  (Other): 8  (Other):555  (c)cd1 :241
                NA's :262  NA's  :2375  NA's  : 1  (Other):585
```

Looking at the survey results, we basically want to count the $c/(c+i)$ (proportion of correct answers) separately for each subject heading (denoted by shnum), for two different kinds of orderings.

```
# let's keep only the attempted SH descriptions
justattempted = surveyresults[(!is.na(surveyresults$correct))&(surveyresults$correct != -1),]

# and look at a summary of what is left
summary(justattempted)

# now we will average over all people who described the same subject heading in the same order
bothorderings = aggregate((justattempted$correct == "c"),
list(order=justattempted$order,sh=justattempted$shnum),FUN=mean)

# paired t-test
> t.test(x ~ order,paired=T,data=bothorderings)
```

Paired t-test

```
data: x by order
t = 0.7796, df = 23, p-value = 0.4436
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.03249104  0.07178732
sample estimates:
mean of the differences
 0.01964814
```

```

#unpaired t-test (WRONG)
> t.test(x ~ order,data=bothorderings)

Welch Two Sample t-test

data: x by order
t = 0.3432, df = 45.982, p-value = 0.733
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0955977  0.1348940
sample estimates:
mean in group o mean in group r
      0.3677402      0.3480920

```

Notice that neither the paired t-test nor the regular one (which is incorrectly applied here), give us significant differences in people's average ability to interpret the subject heading correctly. But the paired t-test does give a lower p-value and a narrower confidence interval, which shows that it is superior. Why is this? Well, there is a lot of variation in the interpretation difficulty of each subject heading. The paired test keeps the variation due to question difficulty separate from the variation due to the two different ways of presenting the subject headings.

I've used the `aggregate()` command to calculate the proportion of people who answered the question correctly in that order. I did this by creating a TRUE/FALSE vector with (`justattempted$correct == "c"`) and grouping by order and subject heading, and then taking the mean. `mean()` appears to treat boolean (TRUE/FALSE) vectors as 0/1 vectors, which means that we can average them.

In any case, since standardizing headings to be in a prescribed (recommended) order did not seem to impact people's ability (and inability) to interpret subject headings, one of the recommendations resulting from this study was to standardize.

2 two-way ANOVA

We may wish to consider two variables (factors) simultaneously, and for this we would do a two-way ANOVA. The example I am using here is made up (and should be familiar from lecture). Recently, there was a NYT article in which it was mentioned that boys and girls may do better when taught in separate classrooms, in part because the conditions could be set to fit each gender separately. An argument was made that boys are more comfortable with the thermostat set to a lower temperature in the classroom. We are going to test this assumption (but with some fake data of course)

```

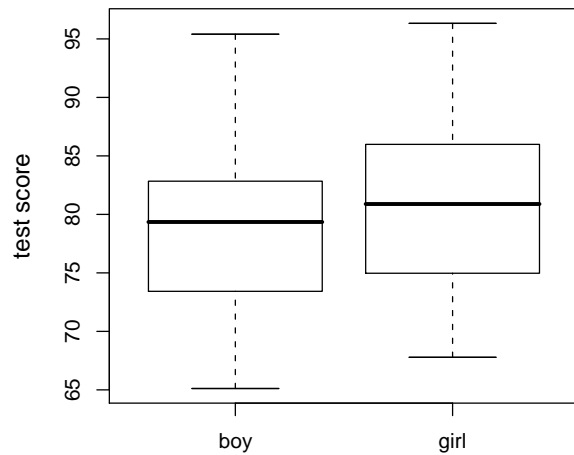
# have 90 different observation, 30 at each
# of the following temperatures: 68, 72, 78
temperature = c(rep(68,30),rep(72,30),rep(78,30))

# we have two different genders boys and girls
gender = rep(c(rep("boy",15),rep("girl",15)),3)

# we translate this to a numerical value
# just to generate some data
gendernum = rep(c(rep(-1,15),rep(1,15)),3)

# now we're ready to create some fake
# data. the kids are going to average a
# score of 80, but the distribution is normal,

```



```

# with a standard deviation of 5.
# In addition, the boys are going to score
# better when it's colder, for the girls it's
# the opposite

testscore = 80 + rnorm(90,0,5) + (temperature - 72)*gendernum

# first we should see no difference in average scores
# for boys and girls when the temperature is not taken
# into account

boxplot(testscore ~ gender,ylab="test score")

# similarly, the t test should not be able to pick up anything
t.test(testscore ~ gender)
> t.test(testscore ~ gender)

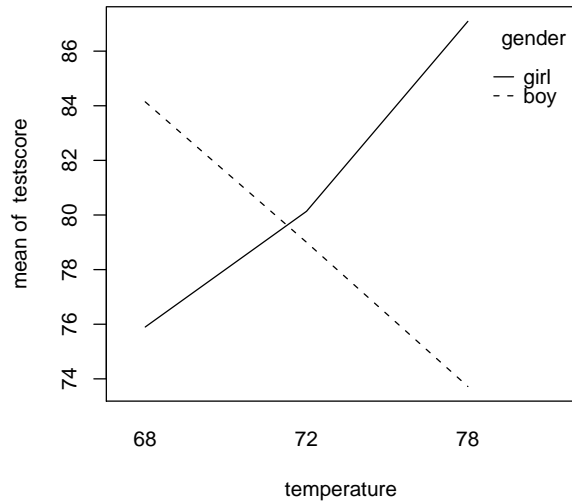
Welch Two Sample t-test

data:  testscore by gender
t = -1.4815, df = 87.997, p-value = 0.1421
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.8887375  0.7128885
sample estimates:
mean in group boy mean in group girl
      78.95415      81.04208

# you can do the same for temperature...

# but now, let's ask for an interaction plot
# this should look familiar from the lecture...
interaction.plot(temperature,gender,testscore)

```



```
# and finally, we can do the ANOVA
anova(lm(testscore ~ temperature*gender))
```

Analysis of Variance Table

Response: testscore

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
temperature	1	3.27	3.27	0.1295	0.71983
gender	1	98.09	98.09	3.8875	0.05186 .
temperature:gender	1	1759.73	1759.73	69.7443	1.027e-12 ***
Residuals	86	2169.88	25.23		

So we discover that even though on average temperature does not seem to matter, and boys and girls score about the same, when we do a **factorial analysis** on both the temperature and gender variables, we discover an **interaction effect**. Whether a child performs better at a given temperature setting actually depends on their gender! (in this made up example only, of course)