AMS 597: Statistical Computing

Pei Fen Kuan
In this section, we consider comparisons among more than two groups parametrically, using analysis of variance.

Let \( x_{ij} \) denote the \( j \)th observation in the \( i \)th group, \( \bar{x}_i \) is the mean of the \( i \)th group, and \( \bar{x}_.. \) is the mean of all observations. We can decompose the observations as

\[
x_{ij} = \bar{x}_.. + (\bar{x}_i - \bar{x}_..) + (x_{ij} - \bar{x}_i)
\]
Informally, this also corresponds to the model

\[ X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \]

in which the hypothesis is that all the group means are same.

Now consider the following sums of squares, known as variation within groups

\[ SSW = \sum_{ij} (x_{ij} - \bar{x}_i)^2 \]

and variation between groups

\[ SSB = \sum_i n_i (\bar{x}_i - \bar{x}_.)^2 \]
One way analysis of variance

- Note that

\[
SSW + SSB = SST = \sum_{ij} (x_{ij} - \bar{x}_{..})^2
\]

- Let MSW = SSW/(N-k), and MSB = SSB/(k-1). We calculate the F-statistics F = MSB/MSW.
One-way analysis of variance

```r
> y1 <- c(18.2, 20.1, 17.6, 16.8, 18.8, 19.3, 19.1)
> y2 <- c(17.4, 18.7, 19.1, 16.4, 15.2, 18.4)
> y3 <- c(15.2, 18.8, 17.7, 16.5, 15.9, 17.1, 16.3)
> y <- c(y1, y2, y3)
> n <- c(7, 6, 7)
> group <- c(rep(1,7), rep(2,6), rep(3,7))
> ydata <- data.frame(y=y, group=factor(group))
> fit <- lm(y~group, data=ydata)
> anova(fit)
```

Analysis of Variance Table

```
Response: y

  Df Sum Sq Mean Sq  F value    Pr(>F)
group      2 11.063  5.5315 3.43960  0.05567 .
Residuals 17 27.339  1.6082
```

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```r
> summary(aov(y~group, data=ydata))
```

```
Df Sum Sq Mean Sq  F value    Pr(>F)
group      2 11.062  5.5312 3.43957  0.05557 .
Residuals 17 27.339  1.6081
```

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Consider the model which decomposes observations into a general level, a row effect, a column effect and a noise term.

\[ X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \]

in which \( \sum \alpha_i = 0 \) and \( \sum \beta_j = 0 \). Let \( x_{ij} \) denote the \( j \)th observation in the \( i \)th row, \( \bar{x}_i \) is the mean of the \( i \)th row, and \( \bar{x}_\cdot \) is the overall mean.
Two way ANOVA

```r
> heart.rate <- data.frame(hr =
c(96,110,89,95,128,100,72,79,100,92,106,86,78,124,98,68,75,106,86,108,85,78,118,100,67,74,+
+ 104,92,114,83,83,118,94,71,74,102),
+ subj=gl(9,1,36),
+ time=gl(4,9,36,labels=c(0,30,60,120)))
> str(heart.rate)
'data.frame': 36 obs. of 3 variables:
$ hr : num 96 110 89 95 128 100 72 79 100 92 ... 
$ subj: Factor w/ 9 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8 9 1 ... 
$ time: Factor w/ 4 levels "0","30","60",...: 1 1 1 1 1 1 1 1 1 2 ... 
```

The `gl` (generate levels) function is specially designed for generating patterned factors for balanced experimental designs.
Two way ANOVA

- It has three arguments: the number of levels, the block length (how many times each level should repeat), and the total length of the result. The two patterns in the data frame are thus

```r
> gl(9,1,36)
[1] 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9 1 2 3 4
[32] 5 6 7 8 9
Levels: 1 2 3 4 5 6 7 8 9
> gl(4,9,36,labels=c(0,30,60,120))
[1] 0 0 0 0 0 0 0 0 0 0 30 30 30 30 30 30 30
[16] 30 30 30 60 60 60 60 60 60 60 60 60 120 120 120 120
[31] 120 120 120 120 120 120
Levels: 0 30 60 120
```
Two way ANOVA

> attach(heart.rate)
> anova(lm(hr~subj+time))

Analysis of Variance Table

Response: hr

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>subj</td>
<td>8</td>
<td>8966.6</td>
<td>1120.82</td>
<td>90.6391</td>
</tr>
<tr>
<td>time</td>
<td>3</td>
<td>151.0</td>
<td>50.32</td>
<td>4.0696</td>
</tr>
<tr>
<td>Residuals</td>
<td>24</td>
<td>296.8</td>
<td>12.37</td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Two way ANOVA

- Discussion on type I, II and III sum of squares (SS)
- Suppose $Y = A + B + A*B$
- Type I SS:
  - $A: SS(A)$
  - $B: SS(B | A)$
  - $AB: SS(AB | A,B)$
Two way ANOVA

- **Type II SS:**
  - A: SS(A | B)
  - B: SS(B | A)
  - AB: SS(AB | A,B)

- **Type III SS:**
  - A: SS(A | B, AB)
  - B: SS(B | A, AB)
  - AB: SS(AB | A,B)
The variation between and within groups for a one-way analysis of variance generalizes to model variation

\[ SS_{model} = \sum_i (\hat{y}_i - \bar{y})^2 \]

and residual variation

\[ SS_{resid} = \sum_i (y_i - \hat{y}_i)^2 \]

which partition the total variation

\[ SS_{total} = \sum_i (y_i - \bar{y})^2 \]
This applies only when the model contains an intercept. The role of the group means in the one-way classification is taken over by the fitted values in the more general linear model.
ANOVA table in regression analysis

```r
> logret <- read.table("http://www.ams.sunysb.edu/~pfkuan/Teaching/AMS597/Data/d_logret_6stocks.txt", header=T)
> attach(logret)
> fit1 <- lm(Pfizer~Intel)
> anova(fit1)
Analysis of Variance Table

Response: Pfizer

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel</td>
<td>1</td>
<td>0.000154</td>
<td>0.00015441</td>
<td>0.2865</td>
<td>0.5944</td>
</tr>
<tr>
<td>Residuals</td>
<td>62</td>
<td>0.033409</td>
<td>0.00053885</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

> fit2 <- lm(Pfizer~Intel+AmerExp)
> anova(fit2)
Analysis of Variance Table

Response: Pfizer

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel</td>
<td>1</td>
<td>0.000154</td>
<td>0.00015441</td>
<td>0.2942</td>
<td>0.5895</td>
</tr>
<tr>
<td>AmerExp</td>
<td>1</td>
<td>0.001396</td>
<td>0.00139571</td>
<td>2.6595</td>
<td>0.1081</td>
</tr>
<tr>
<td>Residuals</td>
<td>61</td>
<td>0.032013</td>
<td>0.00052480</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

> detach(logret)
```
ANOVA table in regression analysis

```r
> SST <- sum((Pfizer-mean(Pfizer))^2)
> SSM <- sum((fit1$fitted-mean(Pfizer))^2)
> SSR <- sum((Pfizer-fit1$fitted)^2)
> SSM+SSR
[1] 0.03356306
> SST
[1] 0.03356306
```
Tests of single proportions are generally based on the binomial distribution with size parameter N and probability parameter p.

For large sample sizes, this can be well approximated by a normal distribution with mean Np and variance Np(1 − p).

As a rule of thumb, the approximation is satisfactory when the expected numbers of “successes” and “failures” are both larger than 5.
Denoting the observed number of “successes” by \( x \), the test for the hypothesis \( p = p_0 \) that can be based on

\[
u = \frac{x - Np_0}{\sqrt{Np_0(1 - p_0)}}\]

which has an approximate normal distribution with mean zero and standard deviation 1.
We consider an example (Altman, 1991, p. 230) where 39 of 215 randomly chosen patients are observed to have asthma and one wants to test the hypothesis that the probability of a “random patient” having asthma is 0.15. This can be done using `prop.test`:
> prop.test(39, 215, 0.15)

1-sample proportions test with continuity correction

data:  39 out of 215, null probability 0.15
X-squared = 1.425, df = 1, p-value = 0.2326
alternative hypothesis: true p is not equal to 0.15
95 percent confidence interval:
  0.1335937 0.2408799
sample estimates:
   p
0.1813953

> binom.test(39, 215, 0.15)

Exact binomial test

data:  39 and 215
number of successes = 39, number of trials = 215, p-value = 0.2135
alternative hypothesis: true probability of success is not equal to 0.15
95 percent confidence interval:
  0.1322842 0.2395223
sample estimates:
probability of success
  0.1813953
Two independent proportions

- The function `prop.test` can also be used to compare two or more proportions. For that purpose, the arguments should be given as two vectors, where the first contains the number of positive outcomes and the second the total number for each group.
The theory is similar to that for a single proportion. Consider the difference in the two proportions $d = x_1/N_1 - x_2/N_2$, which will be approximately normally distributed with mean zero and variance $V(p) = (1/N_1 + 1/N_2) \times p(1 - p)$ if the counts are binomially distributed with the same $p$ parameter.
So to test the hypothesis that $p_1 = p_2$, plug the common estimate $\hat{p} = (x_1 + x_2)/(n_1 + n_2)$ into the variance formula and the test statistic is $d/\sqrt{V(\hat{p})}$, which follows a standard normal distribution approximately. See the following example for illustration.
Two independent proportions

> lewitt.machin.success <- c(9,4)
> lewitt.machin.total <- c(12,13)
> prop.test(l ewitt.machin.success, lewitt.machin.total)

2-sample test for equality of proportions with continuity correction

data: lewitt.machin.success out of lewitt.machin.total
X-squared = 3.2793, df = 1, p-value = 0.07016
alternative hypothesis: two.sided
95 percent confidence interval:
 0.01151032 0.87310506
sample estimates:
  prop 1  prop 2
0.7500000 0.3076923

The confidence interval given is for the difference in proportions.
For the analysis of tables with more than two classes on both sides, you can use `chisq.test` or `fisher.test`, although you should note that the latter can be very computationally demanding if the cell counts are large and there are more than two rows or columns.
r x c tables

An $r \times c$ table looks like this:

\[
\begin{array}{cccc|c}
  n_{11} & n_{12} & \cdots & n_{1c} & n_{1.} \\
  n_{21} & n_{22} & \cdots & n_{2c} & n_{2.} \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  n_{r1} & n_{r2} & \cdots & n_{rc} & n_{r.} \\
  \hline
  n_{.1} & n_{.2} & \cdots & n_{.c} & n_{..}
\end{array}
\]
Such a table can arise from several different sampling plans, and the notion of “no relation between rows and columns” is correspondingly different. The total in each row might be fixed in advance, and you would be interested in testing whether the distribution over columns is the same for each row, or vice versa if the column totals were fixed (Test of homogeneity)
It might also be the case that only the total number is chosen and the individuals are grouped randomly according to the row and column criteria. In the latter case, you would be interested in testing the hypothesis of statistical independence, that the probability of an individual falling into the $ij$th cell is the product $p_i p_j$ of the marginal probabilities. However, the analysis of the table turns out to be the same in all cases (Test of independence)
If there is no relation between rows and columns, then you would expect to have the following cell values:

$$E_{ij} = \frac{n_{i.} \times n_{.j}}{n_{..}}$$

This can be interpreted as distributing each row total according to the proportions in each column (or vice versa) or as distributing the grand total according to the products of the row and column proportions.
The test statistic

\[ X^2 = \sum \frac{(O - E)^2}{E} \]

has an approximate chi-squared distribution with \((r - 1) \times (c - 1)\) degrees of freedom. Here the sum is over the entire table and the \(ij\) indices have been omitted. \(O\) denotes the observed values and \(E\) the expected values as described above.
> caff.marital <- matrix(c(652, 1537, 598, 242, 36, 46, 38, 21, 218, 327, + 106, 67), nrow=3, byrow=T)
> colnames(caff.marital) <- c("0","1-150","151-300",">300")
> rownames(caff.marital) <- c("Married","Prev.married","Single")
> caff.marital

     0 1-150 151-300 >300
Married  652  1537    598   242
Prev.married  36   46     38    21
Single     218  327    106    67
> chisq.test(caff.marital)

    Pearson's Chi-squared test

data:  caff.marital
X-squared = 51.656, df = 6, p-value = 2.187e-09
The test is highly significant, so we can safely conclude that the data contradict the hypothesis of independence. However, you would generally also like to know the nature of the deviations. To that end, you can look at some extra components of the return value of chisq.test().
```r
> chisq.test(caff.marital)$expected
          0 1-150 151-300 >300
Married    705.83179 1488.01183 578.06533 257.09105
Prev.married 32.85648   69.26698  26.90895   11.96759
Single     167.31173  352.72119 137.02572   60.94136
> chisq.test(caff.marital)$observed
          0 1-150 151-300 >300
Married     652   1537    598    242
Prev.married  36     46    38     21
Single       218    327    106     67
> E <- chisq.test(caff.marital)$expected
> O <- chisq.test(caff.marital)$observed
> (O-E)^2/E
          0  1-150  151-300 >300
Married    4.1055981 1.612783 0.6874502 0.8858331
Prev.married 0.3007537 7.815444 4.5713926 6.8171090
Single     15.3563704 1.875645 7.0249243 0.6023355
```