Chapter 11 - Lecture 1 Single Factor ANOVA

Yuan Huang

April 5, 2013

Chapter 9 : hypothesis testing for one population mean.

Chapter 10: hypothesis testing for two population means.

What comes next?

Chapter 9 : hypothesis testing for one population mean.

Chapter 10: hypothesis testing for two population means.

What comes next?

3 or more population means!

Examples of multiple population comparisons:

- Compare safety factors for multiple makers of cars.
- Compare effect of the same drug in different doses.
- Compare survival time for different treatments for cancers.
- Compare the yield of crop with different pesticides.

Father of Modern statistics - Sir R.A. Fisher

- ANOVA 1918, 1919 (Rothamsted Experimental Station (studies in crop variation)).

Definitions

- Factor: The characteristic that labels the populations.
- Levels: The populations that are referred to.

Example: An experiment to study the effects of five different brands of gasoline on automobile engine operating efficiency

- 5 populations;
- Factor: brands of gasoline (single factor);
- Levels: Exxon, Sheetz, Snapper, Shell, X-mobile.

Example : An experiment to study the effect of sunshine and water in growing corns. Suppose the sunshine can be: intense, weak, bare; amount of water can be: much, little.

- $2 \times 3 = 6$ populations;
- Factors: sunshine and water (two factors);
- Levels of factors: intense, weak and bare for sunshine; much and little for water.

When the response is numerical and factor is categorical (finite number of levels), we could apply the analysis of variance **ANOVA**. Depending on how many factors you have in your study (to label your population), comparisons can be classified as:

- Single-factor study (One-way ANOVA);
- Two-factor study (Two-way ANOVA).

In this lecture, we are introducing One-way ANOVA.

Analysis of Variance

We assume I populations

$$\mu_1 = \text{the mean of population 1;}$$
$$\ldots$$
$$\mu_I = \text{the mean of population } I.$$

Hypothesis:

$$H_0 : \mu_1 = \ldots = \mu_I \text{ vs } H_1 : \text{at least two } \mu_i's \text{ differ.}$$

There are three assumptions:

- Each sample has the same size, denoted by J - the data is balanced.
- The populations are normally distributed with mean μ_i.
- Equal variance: $\sigma_1^2 = \ldots = \sigma_I^2$.

Note: To verify equality of variances, there is a formal test called the Levene test. A rule of thumb that one can use is that the largest standard deviation is not larger than two times the smaller.

Example

Let me first give some numbers to help understand the notations will be introduced.

- Assume I am teaching on three different sections of stat 200 and I am giving them a test. I want to test if the true averages on the test for all classes are equal. I am selecting a sample of 5 students from each class.
- Class 1: 70, 50, 100, 100, 70
- Class 2: 60, 85, 65, 100, 30
- Class 3: 80, 50, 90, 75, 85

Sample Mean and Grand Mean

X_{ij}: denote the j^{th} observation in the i^{th} sample.

The **individual sample means** :

$$\bar{X}_{i.} = \frac{\sum_{j=1}^{J} X_{ij}}{J}$$

The **grand mean** is the pooled sample mean

$$\bar{X}_{..} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} X_{ij}}{IJ}$$

Sample Variance

Sample variance:

$$S_i^2 = \frac{\sum_{j=1}^{J} (X_{ij} - \bar{X}_{i.})^2}{J - 1}$$

Idea

Hypothesis:

$$H_0 : \mu_1 = \ldots = \mu_I \text{ vs } \; H_1 : \text{at least two } \; \mu_i' s \; \text{ differ.}$$

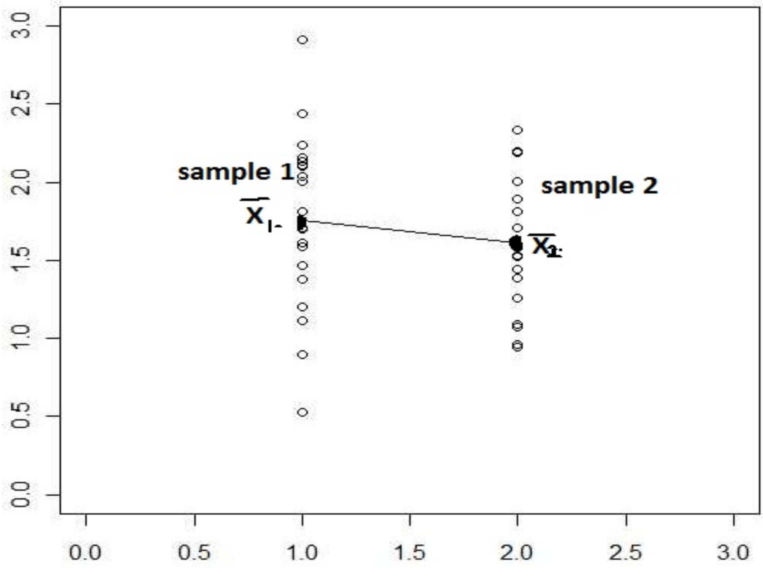
- In order for the null hypothesis to be true, we expect the sample means $\bar{x}_{i.}$ to be as close to the grand mean $\bar{x}_{..}$ as possible.
- How to define close?

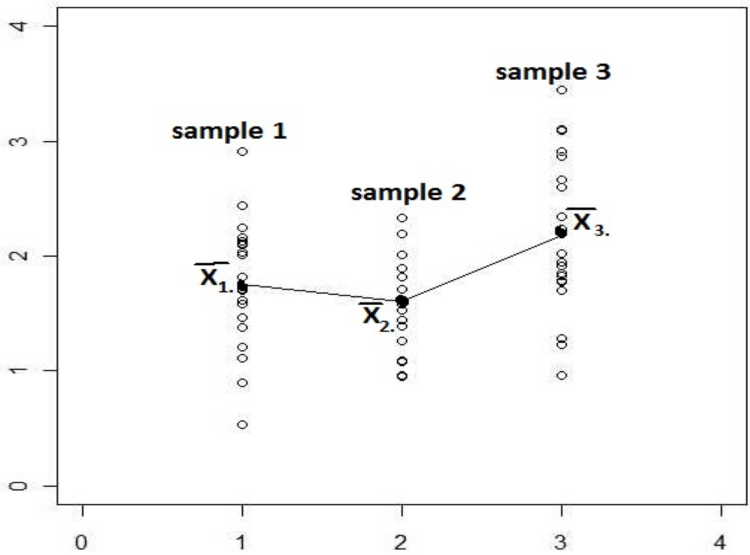
What we did in the last chapter:
For example in the pooled two-sample t test, if $m = n$;

$$T = \frac{\bar{X}_{1\cdot} - \bar{X}_{2\cdot} - 0}{\sqrt{S_p^2 \times 2/n}} \sim t_{2n-2}$$

$$T^2 = \frac{(\bar{X}_{1\cdot} - \bar{X}_{\cdot\cdot})^2 + (\bar{X}_{2\cdot} - \bar{X}_{\cdot\cdot})^2}{S_p^2/n} \sim F_{1,2n-2}$$

This test can be easily generalized.


sample 1
$\overline{X}_{1\cdot}$
sample 2
$\overline{X}_{2\cdot}$
0.0
0.5
1.0
1.5
2.0
2.5
3.0


sample 1
sample 2
sample 3
$\overline{X}_{1.}$
$\overline{X}_{2.}$
$\overline{X}_{3.}$
0
1
2
3
4
0
1
2
3
4

We expect our test can be based on the following statistic:

$$\frac{(\bar{X}_{1\cdot} - \bar{X}_{\cdot\cdot})^2 + (\bar{X}_{2\cdot} - \bar{X}_{\cdot\cdot})^2 + (\bar{X}_{3\cdot} - \bar{X}_{\cdot\cdot})^2}{S_p^2/n}$$

More generally, if the factor has I levels, it will be

$$\frac{(\bar{X}_{1\cdot} - \bar{X}_{\cdot\cdot})^2 + \ldots + (\bar{X}_{I\cdot} - \bar{X}_{\cdot\cdot})^2}{S_p^2/n}$$

Error Sum of Square *SSE*

Definition: error sum of squares SSE:

$$SSE = \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{X}_{ij} - \bar{X}_{i.})^2 = (J-1)\sum_{i=1}^{I} S_i^2$$

- Distribution of SSE:

$$\frac{SSE}{\sigma^2} \sim \chi^2_{I(J-1)}$$

Treatment sum of square *SSTr*

Definition: treatment sum of squares SSTr:

$$SSTr = J \sum_{i=1}^{I} (\bar{X}_{i.} - \bar{X}_{..})^2$$

- Distribution of SSTr: (If H_0 is true)

$$\frac{SSTr}{\sigma^2} \sim \chi^2_{I-1}$$

Definition: Sum of Squares Total (SST):

$$SST = \sum_i \sum_j (X_{ij} - \bar{X}_{..})^2$$

Mathematically, under the assumption of balanced data, $SST = SSTr + SSE$, called **the fundamental ANOVA identity**.

Recall: $H_0 : \mu_1 = \ldots = \mu_I$. Denote $\sigma^2 = \sigma_1^2 = \ldots = \sigma_I^2$

Theorem: Under all the assumptions (balanced data, normal distribution, equal variance), we have

- $SSE/\sigma^2 \sim \chi^2_{I(J-1)}$ no matter H_0 is true or not;
- $SSTr/\sigma^2 \sim \chi^2_{I-1}$ if and only if H_0 is true;
- $SSTr$ and SSE are independent random variables.

where

$$
\begin{aligned}
SSE &= \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{X}_{ij} - \bar{X}_{i.})^2 = (J-1)\sum_{i=1}^{I} S_i^2 \\
SSTr &= J\sum_{i=1}^{I}(\bar{X}_{i.} - \bar{X}_{..})^2
\end{aligned}
$$

Comments:

- This theorem can be easily verified in the two-sample case - it's equivalent to the pooled two-sample t test.
- An immediate result is:

$$\frac{SSTr/[\sigma^2(I-1)]}{SSE/[\sigma^2 I(J-1)]} \sim F_{I-1,I(J-1)} \text{ under } H_0$$

- If H_0 is not true, the F value tends to be large.

Test

The F test:

$$F = \frac{MSTr}{MSE} = \frac{SSTr/[\sigma^2(I-1)]}{SSE/[\sigma^2 I(J-1)]}$$

$F \sim F_{I-1,I(J-1)}$ under H_0. We reject H_0 at level α whenever $F > F_{\alpha,I-1,I(J-1)}$.

Definition:

- the **mean square for treatments** is $MSTr = SSTr/(I-1)$.
- the **mean square for error** is $MSE = SSE/[I(J-1)]$.

Comments:

1. Mean square = Sum of square / corresponding degree of freedom
2.
   1. MSE is the pooled sample estimator of σ^2
   $$E\left(\frac{SSE}{I(J-1)}\right) = E(MSE) = \sigma^2$$
   2. Under H_0, MSTr is also estimating σ^2; if H_0 is true:
   $$E\left(\frac{SSTr}{I-1)}\right) = E(MSTr) = \sigma^2$$
   3. Under H_1, MSTr is estimating a larger parameter.

ANOVA Table

- All the previous results can be summarized in the following Table:

Table: ANOVA TABLE

Source of variation	df	Sum of Squares	Mean Squares	F
Treatments	$I-1$	$SSTr$	$MSTr$	$MSTr/MSE$
Error	$I(J-1)$	SSE	MSE	
Total	$IJ-1$	SST		

Summarization

If the means of ≥ 3 populations are compared:

- check assumptions:
  1. Independence between samples (for one-way ANOVA): based on how the data is collected.
  2. Balanced Data: each sample has the same size J;
  3. Normality Assumption: If J is large, use normal probability plot in each sample; otherwise draw this plot for pooled $x_{ij} - \bar{x}_{i\cdot}$;
  4. Equal Variance
- Fill out (one-way) ANOVA table.
- Make decision based on F-test.

The most important thing in practice (if a stat software is available) is:

1. to know when ANOVA can be used;
2. to know how to read the ANOVA table from the output.

From Minitab:

```
Analysis of Variance
Source     DF        SS        MS        F        P
Factor      3   4703.19   1567.73   206.72    0.000
Error      16    121.34      7.58
Total      19   4824.53
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
Level       N      Mean     StDev  --+---------+---------+---------+----
Feed 1      5     60.68      3.03   (*-)
Feed 2      5     69.24      2.96        (-*-)
Feed 3      5    100.34      2.16                             (-*-)
Feed 4      5     86.38      2.78                     (-*)
                                   --+---------+---------+---------+----
Pooled StDev =     2.75              60        75        90       105
```

From R:

Source	*DF*	*Sum of squares*	*Mean squares*	*F ratio*	*F probability*
Between	2	1,961.63	980.81	1.1868	0.3085
Within	131	10,8261.69	826.42		
Total	133	11,0223.32			

Group	*N*	*Mean per cent*	*STD error*
Electronic products	97	42.47	2.92
Recreational equipment	26	24.04	5.62
Appliances	11	33.63	8.58
Total	134	40.12	2.48

Note:
Bartlett's Box $f = 0.001$ $p = 0.999$