§2 Hand-out 2

Dr. M. P. M. M. McLoughlin

Revised 2018

3. Fundamentals

3.1. **Preliminaries.** Suppose we can produce a random sample of weights of 10 year-olds (in lbs.) from China, the United States, and Russia. We wish to compare the weights of the children, by country, to determine if they differ significantly and can such differences be attributed to chance or not.

Let μ_C be the mean weight of Chinese 10-yr. olds; σ_C^2 be the variance of weight of Chinese 10-yr. olds; μ_A be the mean weight of American 10-yr. olds; σ_A^2 be the variance of weight of American 10-yr. olds; μ_R be the mean weight of Russian 10-yr. olds; and, σ_R^2 be the variance of weight of Russian 10-yr. olds.

Someone unfamiliar with proper design of experiments would possibly do the following:

They set α; etc. and will do three Independent T-tests with the first $H_0 : \mu_C = \mu_A$ versus $H_A : \mu_C \neq \mu_A$; then second testing the hypothesis the weights of Chinese v. Russian does not differ; and, finally testing the hypothesis the weights of American v. Russian does not differ.

Said person did the experiment incorrectly. They actually increased type-I error probabilities.

The correct design is the one-way Analysis of Variance (ANOVA) design which is a generalisation of the T-test for multiple groups. It is stated as:
$H_0 : \mu_C = \mu_A = \mu_R$ versus H_A : at least one pair of these means differs significantly from at least one other.

4. One-way Analysis of Variance (ANOVA)

Continuing with the motivating problem from section 1 (above) consider a researcher wishes to compare the weights of the children, by country, to determine if they differ significantly and can such differences be attributed to chance or not.

She sets $\alpha = 0.05$; assumes the population(s) are normally distributed with μ_C; μ_A; μ_R; σ_C; σ_A; and σ_R all existing such that the standard deviations are equal and not zero. Assume the observations (that will be done) are independent identically distributed (i.i.d.) and there is no error of measurement in the collection or recording of the data.

$H_0 : \mu_C = \mu_A = \mu_R$ versus H_A : at least one pair of theses means differs significantly from at least one other.

Let $\bar{X}_C$ be the sample mean weight of Chinese 10-yr. olds; S_C^2 be the sample variance of weight of Chinese 10-yr. olds; $\bar{X}_A$ be the sample mean weight of American 10-yr. olds; S_A^2 be the sample variance of weight of American 10-yr. olds; $\bar{X}_R$ be the sample mean weight of Russian 10-yr. olds;

and, S_R^2 be the sample variance of weight of Russian 10-yr. olds.

The lettering is cumbersome so we shall switch to and the Chinese group is group 1, the Americans group 2; and the Russians group 3. So, $H_0 : \mu_1 = \mu_2 = \mu_3$ versus H_A : at least one pair of theses means differs significantly from at least one other.

The data is as follows:

China: $72, 58, 74, 66, 70$.

USA: $76, 85, 82, 80, 77$.

Russia: $77, 81, 71, 76, 80$.

Notationally we do as follows:

China: $x_{11}, x_{12}, x_{13}, x_{14}, x_{15}$.

USA: $x_{21}, x_{22}, x_{23}, x_{24}, x_{25}$.

Russia: $x_{31}, x_{32}, x_{33}, x_{34}, x_{35}$.

Let $i \in \mathbb{N}_3$ and $j \in \mathbb{N}_5$. Each observation can be thought of as follows:

$X_{ij} \sim N(\mu_i, \sigma_i)$.

$x_{ij} = \mu_i + \varepsilon_{ij}$ where ϵ_{ij} is the **error** or individual differences within the group. For ANOVA, it is assumed that $\epsilon_{ij} \sim N(0, 1)$.

It is assumed under H_0 there is no difference in the means so let μ be the **grand mean**. Each observation can be defined as follows:

$x_{ij} = \mu + \delta_i + \varepsilon_{ij}$ where δ_i is the **treatment effect** (by group).
Under H_0, $\sum_{i=1}^{3} \delta_i = 0$; therefore, $\mu_i = \mu + \delta_i$ and it must be the case that the mean of the means is the grand mean (in notation: $E[\mu_i] = \mu \quad \forall i \in \mathbb{N}_3$?).

Recall $H_0 : \mu_1 = \mu_2 = \mu_3$ versus H_A : at least one pair of theses means differs significantly from at least one other which is logically equivalent to $H_0 : \delta_i = 0 \quad \forall i \in \mathbb{N}_3$ versus $H_A : \delta_i \neq 0$ for at least one $i \in \mathbb{N}_3$.

In general, let is have a data set of n categories ($n \in \mathbb{N}$) each of size m ($m \in \mathbb{N}$)

Group 1: $x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, \ldots x_{1m}$.

Group 2: $x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, \ldots x_{2m}$.

$\vdots$

Group n: $x_{n1}, x_{n2}, x_{n3}, x_{n4}, x_{n5}, \ldots x_{nm}$.

The **grand mean of the sample** shall be denoted as $\bar{x}_{..}$; it is defined for a sample with n categories and m observations per category ($n \in \mathbb{N}; m \in \mathbb{N}$): $\bar{x}_{..} = \frac{1}{n} \cdot \frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}$

The sample mean of group 1 is denoted as $\bar{x}_{1\cdot}$.
The sample mean of group 2 is denoted as $\bar{x}_{2\cdot}$.

$\vdots$

The sample mean of group n is denoted as $\bar{x}_{n\cdot}$.

And, the sample grand mean is denoted as $\bar{x}_{..}$

For our motivating example from above, it is $\bar{x}_{..} = \frac{1}{3} \cdot \frac{1}{5} \sum_{i=1}^{3} \sum_{j=1}^{5} x_{ij}$.
The analysis is on the total variability of the combined groups data which is

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - \bar{x}_{..})^2.$$

This quantity is referred to as the Sum of Squares Total (SST).

Definition 4.1. Let there be a normal population with existing mean and non-zero variance or n normal populations with existing means and non-zero variance. The Sum of Squares Total (SST) is defined as the quantity

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - \bar{x}_{..})^2.$$

Under H_0 all of this variability is owed to chance. Hence, if the H_0 is false there is something 'going on' due to the differences between the groups (in our example, there is variability due to individuals (obviously) but the H_0 is false there seems to be a difference reflected by being Chinese, Russian, or American).

Theorem 4.1. *Let there be a normal population with existing mean and non-zero variance or n normal populations with existing means and non-zero variance.*
$\sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{i\cdot} - \bar{x}_{..})^2 + \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - \bar{x}_{i\cdot})^2$ *where $\bar{x}_{i\cdot}$ is the mean for group i and $\bar{x}_{..}$ is the grand mean.*

Recall the Sum of Squares Total (SST) is

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij}-\bar{x}_{..})^2$$

Definition 4.2. Let there be a normal population with existing mean and non-zero variance or n normal populations with existing means and equal non-zero variance. The Sum of Squares Between (SSB) groups (also called the sum of squares treatment) is defined as the quantity

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{i\cdot}-\bar{x}_{..})^2$$

Definition 4.3. Let there be a normal population with existing mean and non-zero variance or n normal populations with existing means and equal non-zero variance. The Sum of Squares Within (SSW) group (also called the sum of squares error) is defined as the quantity

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij}-\bar{x}_{i\cdot})^2$$

So, $SST = SSB + SSW$.

SSW measure the chance variation with each group sample

SSB measure the chance variation under the H_0; but, reflects the variation amongst the groups when H_0 is not true.

We did assume a constant fixed variance, σ^2, so we can consider the quantity

$$\frac{\sum_{j=1}^{m}(x_{ij}-\bar{x}_{i\cdot})^2}{\sigma^2}$$

which is a chi-squared random variable with $df = m-1$

$$\frac{SSW}{\sigma^2}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij}-\bar{x}_{i\cdot})^2}{\sigma^2}$$

is a chi-squared random variable with $df = n(m-1)$ Therefore,

$$\frac{SSW}{\sigma^2}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij}-\bar{x}_{i\cdot})^2}{\sigma^2}$$

is a chi-squared random variable with $df = n(m-1)$. Further, it is an estimate of the population variance (it is a $\widehat{\sigma^2}$).

Under H_0 the observations were independent identically distributed normal random variables $X \sim N(\mu, \sigma)$ so the $\bar{X}$ are i.i.d. normal random variables $X \sim N(\mu, \frac{\sigma}{\sqrt{n}})$; thus, considering

$$SSB = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{i\cdot}-\bar{x}_{..})^2$$

we see that

$$m \cdot SSB = m \cdot \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{i\cdot}-\bar{x}_{..})^2$$

which implies

$$\frac{m \cdot SSB}{\sigma^2}=\frac{m \cdot \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{i\cdot}-\bar{x}_{..})^2}{\sigma^2}$$

is a chi-squared random variable with $df = n - 1$. The mean of this $\chi^2_{(n-1)}$ is $n - 1$. Furthermore, $\frac{SSB}{n-1}$ is an estimator of σ^2

Definition 4.4. Let there be a normal population with existing mean and non-zero variance or n normal populations with existing means and equal non-zero variance. The mean squared between (MSB) is defined as $\frac{SSB}{n-1}$.

Definition 4.5. Let there be a normal population with existing mean and non-zero variance or n normal populations with existing means and equal non-zero variance. The mean squared error (MSE) is defined as $\frac{SSW}{n\cdot(m-1)}$

Hence, if H_0 is false then MSB is an estimate of σ^2 with whatever variation there may be amongst the group population means implying that when there is sufficient evidence to reject H_0 at an alpha-level, MSB is 'significantly' greater than MSE.

One must assume the SSB and SSW (MSB and MSE) are independent to be able to consider:

$$\frac{\frac{SSB}{\sigma^2\cdot(n-1)}}{\frac{SSW}{n\cdot(m-1)\cdot\sigma^2}} = \frac{MSB}{MSE}$$

is distributed F (Fisher - Snedecor) with $(n-1)$ and $n\cdot(m-1)$ degrees of freedom. So, at an α level H_0 is rejected provided the computed value of the F-statistic based on the sample is larger than $F_{(\alpha,(n-1),n\cdot(m-1))}$.

If one is doing the calculations by hand - there is a handy (haha) theorem[1] to use:

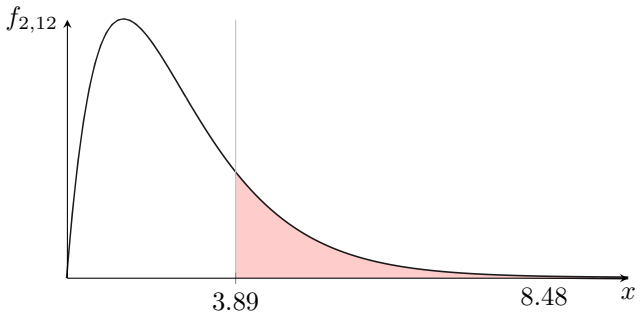
Theorem 4.2. *Let there be a normal population with existing mean and non-zero variance or n normal populations with existing means and non-zero variance.*

$$SST = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij}^2) - \frac{1}{n\cdot m}\cdot T_{\cdot\cdot}^2$$

$$SSB = \frac{1}{n}\cdot\sum_{i=1}^{n}T_{i\cdot}^2 - \frac{1}{n\cdot m}\cdot T_{\cdot\cdot}^2$$

where $T_{i\cdot}$ is the total of the observations for the i^{th} group and $T_{\cdot\cdot}^2$ is the grand total for all the observations.

One computes SSW by subtraction SSB from SST.
Returning to the motivating problem notice:
$F_{(0.05,2,12)} = 3.89$ and we reject H_0 if our computed F is larger.



[1]from Freund, J. *Mathematical Statistics (2nd Ed.) Saddle Brrok, NJ: Pearson.*

$T_{1\cdot} = 385; T_{2\cdot} = 340; ; T_{3\cdot} = 400 \wedge,\ T_{\cdot\cdot} = 1125$. Further, $\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij}^2) = 85041$. $n = 3, m = 5 \implies SST = 85041 - \frac{1}{15}(1125)^2 = 666$ and $SSB = \frac{1}{5}(385^2 + 340^2 + 400^2) - \frac{1}{15}(1125)^2 = 390$ Therefore, $SSW = 276$. Thus, $MSB = \frac{390}{2}$ and $MSE = \frac{276}{12}$. Ergo, $F = \frac{195}{23} \approx 8.48$. Since $8.48 > 3.89$ there is sufficient evidence at the 0.05 - level to reject the null and concluded there seems to be a statistically significant difference in heights (amongst these groups). However, which, where, etc. ? To answer such requires a post-hoc analysis of the groups.

5. Post-Hoc

Let us discuss some of the standard a post-hoc analyses that are typically employed.

1. **The Scheffé method** for investigating all possible contrasts of the means corresponds exactly to the F-test if the F-test rejects the null hypothesis at level α, then there exists at least one contrast which would be rejected using the Scheffé procedure at level α. Therefore, Scheffé provides α level protection against rejecting the null hypothesis when it is true, regardless of how many contrasts of the means are tested.

2. **The Fishers LSD** is the F test, followed by ordinary t-tests among all pairs of means, but only if the F-test rejects the null hypothesis. The F-test provides the overall protection against rejecting H_0 when it is true. The t-tests are each performed at α level and thus likely will reject more than they should, when the F-test rejects.
A simple example may explain this statement: assume there are eight treatment groups, and one treatment has a mean higher than the other seven, which all have the same value, and the F-test will reject H_0. However, when following up with the pairwise t-tests, the $\frac{7\cdot 6}{2} = 21$ pairwise t-tests among the seven means which are all equal, will by chance alone reject at least one pair-wise hypothesis, $H_0 : \mu_j = \mu_k$ for some $j, k \in \mathbb{N}_8$ at the $\alpha = 0.05$ level.

$$T = \frac{\bar{x}_j - \bar{x}_k}{\sqrt{MSE \cdot (\frac{1}{n_j} + \frac{1}{n_k})}}$$

$\forall \quad j, k \in \mathbb{N}_8 \quad \ni \quad j \neq k$ Fisher's LSD is a viable method since it has overall α level protection (and is simple to understand, I suppose).

3. **The Bonferroni method** for g comparisons use α/g instead of α for testing each of the g comparisons. The Bonferroni method does control the family error rate, by performing the pairwise comparison tests using α/g level of significance, where g is the number of pairwise comparisons. Hence, the Bonferroni confidence intervals for differences of the means are 'wide.'

4. **Tukeys Studentised Range** considers the differences among all pairs of means divided by the estimated standard deviation of the mean, and compares them with the tabled critical values provided in handout 3. It is called the studentised range because the denominator of the statistic is an estimated standard deviation; hence, the statistic is studentised like the student t-test. The Tukey procedure assumes all n_i are equal ($i \in \mathbb{N}$). Call it n.

$$q = \frac{\bar{x}_j - \bar{x}_k}{\sqrt{MSE \cdot \frac{1}{n}}}$$

Example 5.1. *Consider* 5 *means, so* $a = 5$*, let* $\alpha = 0.05$*, and the total number of observations be* 35*; so, each group has seven observations and df* $= 30$*. Consider the Tukey studentised range distribution for* $5, 30$ *degrees of freedom, (Handout Gamma), the critical value is* 4.11*. So, when* $q > 4.11$ *for some pair of groups* j, k *then, it is said that those groups means are pair-wise significantly different.*

Consider 5 *means, so* $a = 5$*, let* $\alpha = 0.05$*, and the total number of observations be* 35*; so, each group has seven observations and* $df = 30$*. Consider the Bonferroni approach -* $g = (5 \cdot 4)/2 = 10$ *pair-wise comparisons since* $a = 5$*. Thus, again for* $\alpha = 0.05$ *test consider the t-distribution for* $\alpha/2g = 0.0025$ *and* 30 *df and note a critical value of* 3.03*. So, when* $t > 3.03$ *for some pair of groups* j, k *then, it is said that those groups means are pair-wise significantly different.*

However, to compare with the Tukey Studentised Range statistic, the t - tabled critical value is equivalent to $\sqrt{2}$ *times the t; therefore, a Bonferroni t of* 3.03 *is equivalent to a Tukey q of* $3.03 \cdot \sqrt{2} \approx 4.28$*, which is slightly larger than the* 4.11 *obtained for the Tukey table. Hence, the Bonferroni is more 'conservative' than the Tukey.*

The Tukey procedure is equal samples sizes; if the sample sizes are not equal there is an approximate procedure called the Tukey-Kramer test for unequal n_i*.*