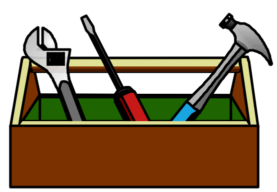
Statistical Techniques II

EXST7015

Analysis of Variance



Design

- **The simplest model for Analysis of Variance (ANOVA) is the CRD, the Completely Randomized Design**
- **This model is also known a "One-way" Analysis of Variance.**
- **Unlike regression, which fits slopes for regression lines and calculates a measure of random variation about those lines, ANOVA fits means and variation about those means.**

Design *(continued)*

- The hypotheses tested are hypotheses about the equality of means
- H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \ \dots \ = \mu_t$
  - Where
    - the μ_i represent means of the levels of some categorical variable
    - "t" is the number of levels in the categorical variable.
- H_1: some μ_i is different

Design *(continued)*

- We will generically refer to the categorical variable as the "treatment" even though it may not actually be an experimenter manipulated effect.
  - The number of treatments will be designated "t".
  - The number of observations within treatments will be designated n for a balanced design (the same number of observations in each treatment), or n_i for an unbalanced design (for i = 1 to t).

Design *(continued)*

- **The assumptions for basic ANOVA are very similar to those of regression.**
  - **The residuals, or deviations of observations within groups should be normally distributed.**
  - **The treatments are independently sampled.**
  - **The variance of each treatment is the same (homogeneous variance).**

ANOVA review

- I am borrowing some material from my EXST7005 notes on t-test and ANOVA. See those notes for a more complete review of the introduction to Analysis of Variance (ANOVA).
- Start with the logic behind ANOVA.

ANOVA review *(continued)*

- **Prior to R. A. Fisher's development of ANOVA, investigators were likely to have used a series of t tests to test among t treatment levels.**
- **What is wrong with that? Recall the Bonferroni adjustment. Each time we do a test we increase the chance of error. To test among 3 treatments we need to do 3 tests, among 4 treatments, 6 tests, 5 treatments are 10 tests, etc.**

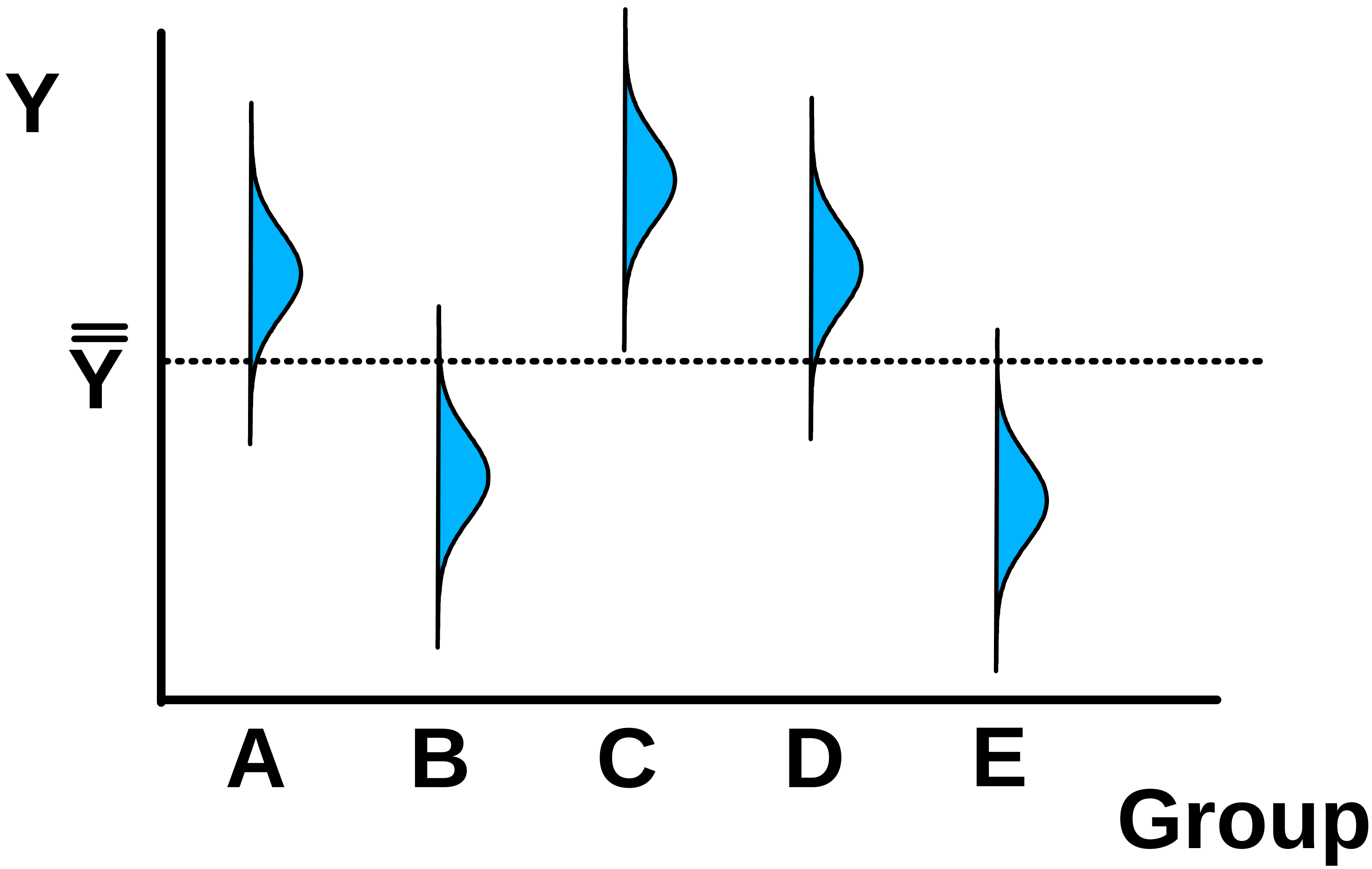
ANOVA review *(continued)*

- **What is needed is ONE test for a difference among all tests with one overall value of α specified by the investigator (usually 0.05).**
- **Fisher's solution was simple, but elegant.**

ANOVA review *(continued)*

- **Suppose we have a treatment with 5 categories or levels. We can calculate a mean and variance for each treatment level. In order to get one really good estimate of variance we can pool the individual variances of the 5 categories (assuming homogeneity of variance).**
- **This pooled variance can be calculated as a weighted mean of the variance (weighted by the degrees of freedom).**

ANOVA review (*continued*)



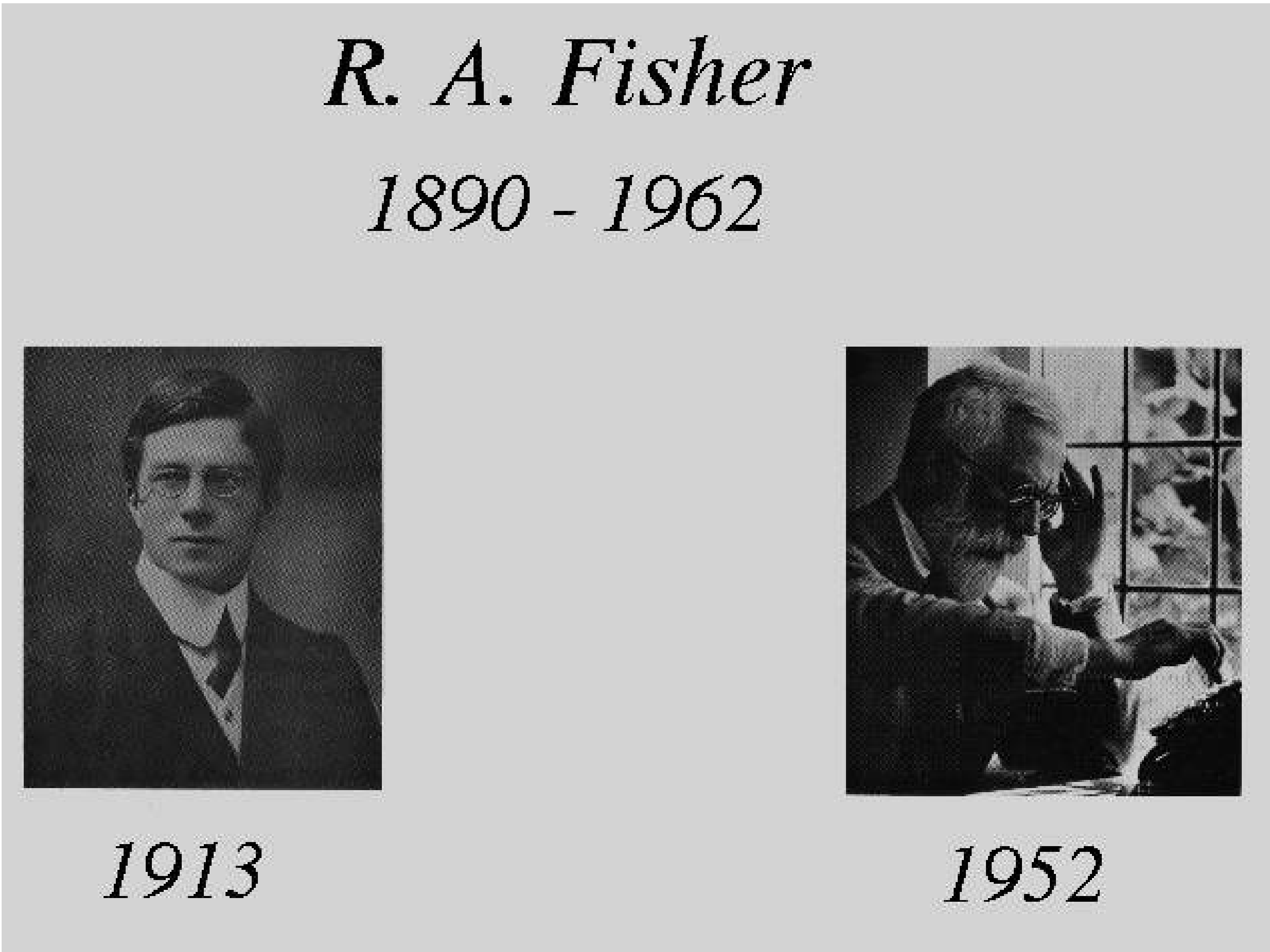
ANOVA review *(continued)*

- **And since $(n_1\text{-}1)S_1^2 = SS_1$, the weighted mean is simply the sum of the SS divided by the sum of the d.f.**

$$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2+(n_3-1)S_3^2+(n_4-1)S_4^2+(n_5-1)S_5^2}{(n_1-1)+(n_2-1)+(n_3-1)+(n_4-1)+(n_5-1)}$$

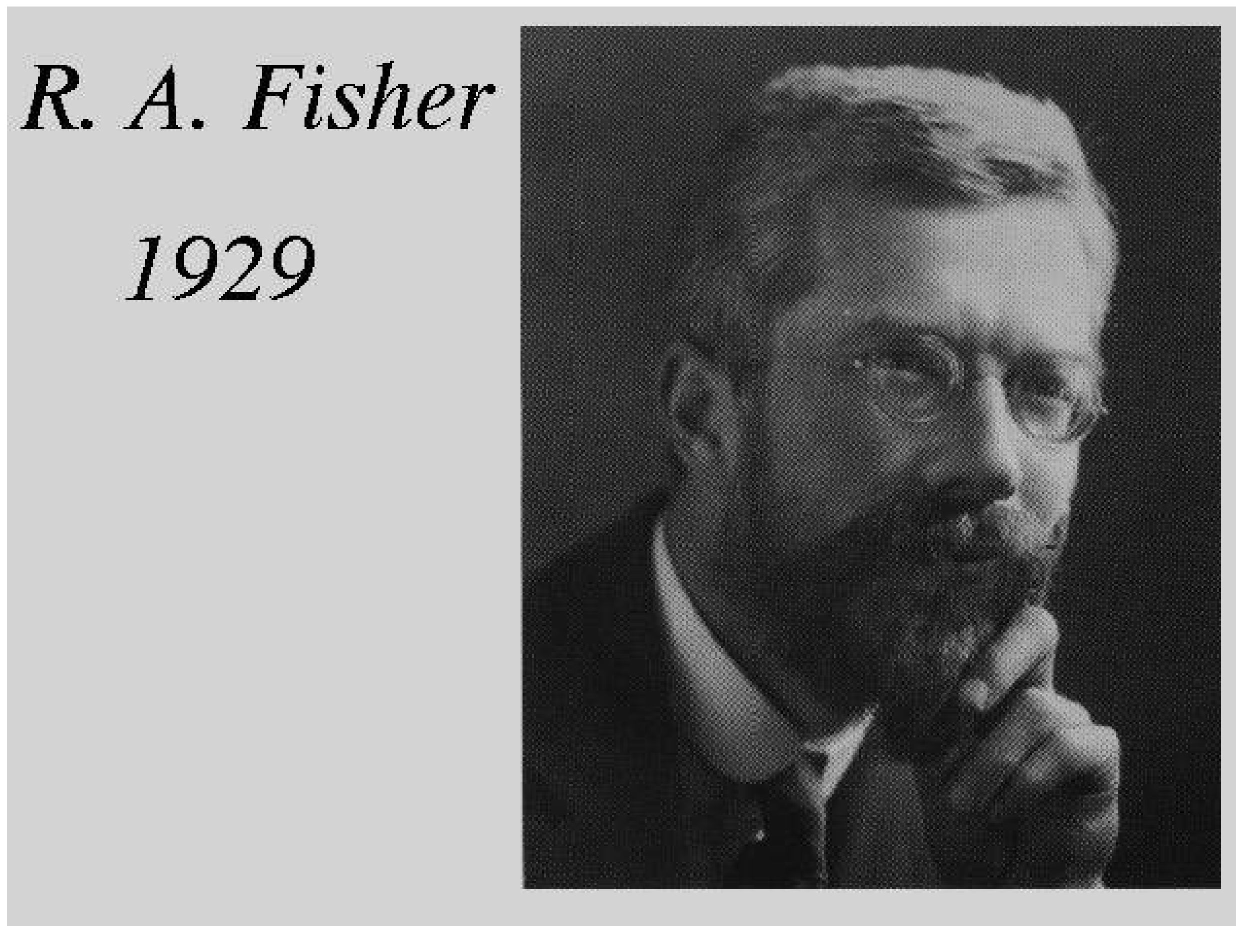
$$S_p^2 = \frac{SS_1+SS_2+SS_3+SS_4+SS_5}{(n_1-1)+(n_2-1)+(n_3-1)+(n_4-1)+(n_5-1)}$$

ANOVA review *(continued)*



ANOVA review *(continued)*

- **So we have one very good estimate of the random variation, or sampling error, S^2.**
- **Then what?**



ANOVA review *(continued)*

- **Now consider the treatments. Why don't they all fall on the overall mean?**
- **Actually, under the null hypothesis, they should, except for some random variation.**
- **So if we estimate that random variation, it should be equal to the same error we already estimated within groups?**

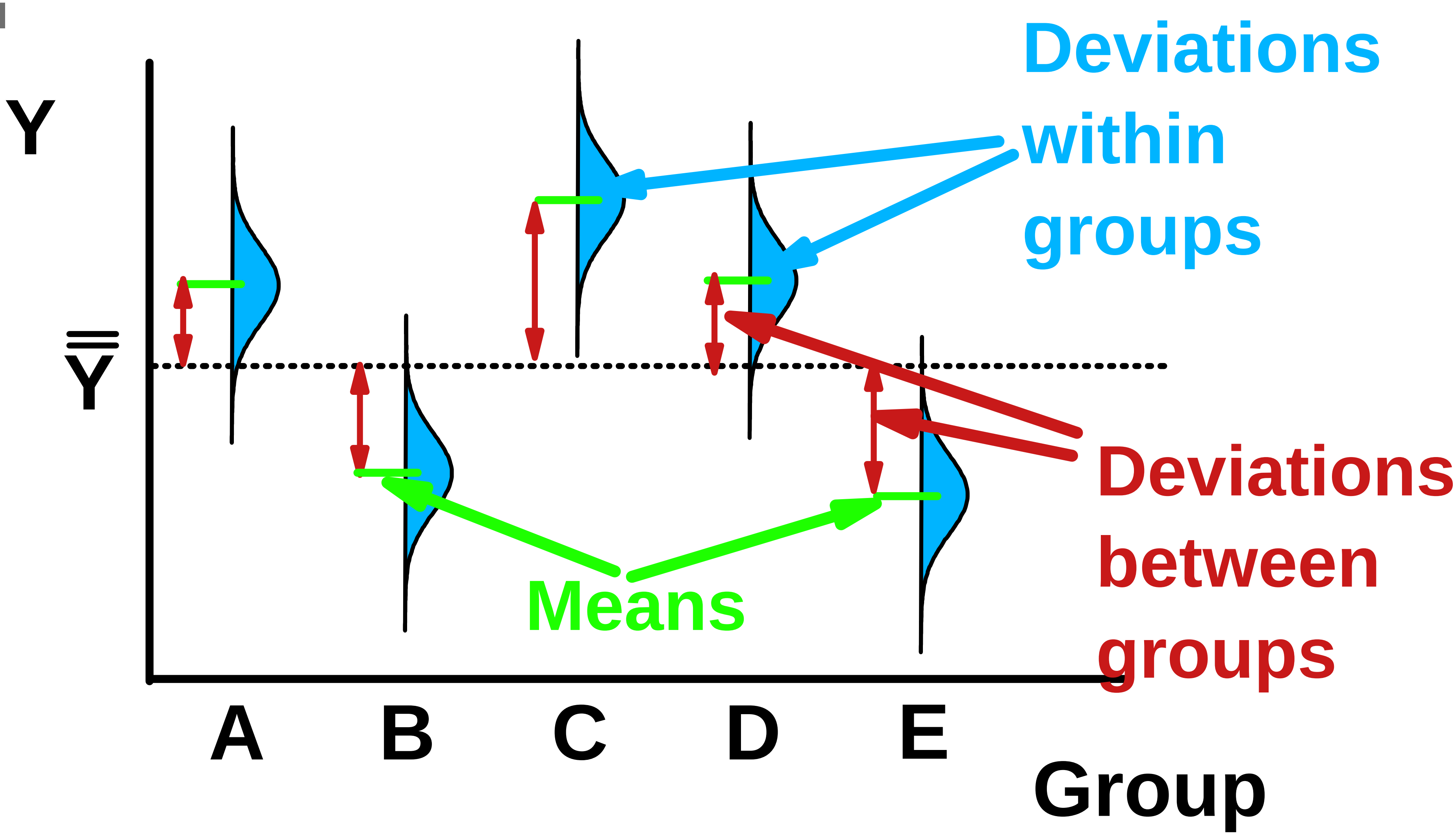
ANOVA review *(continued)*

- **If we estimate a variance with means, we are estimating the variance of means, which is S^2/n. If we multiply this by "n" it should actually be equal to S^2, which we estimated with S^2_p, the pooled variance estimate.**

ANOVA review *(continued)*

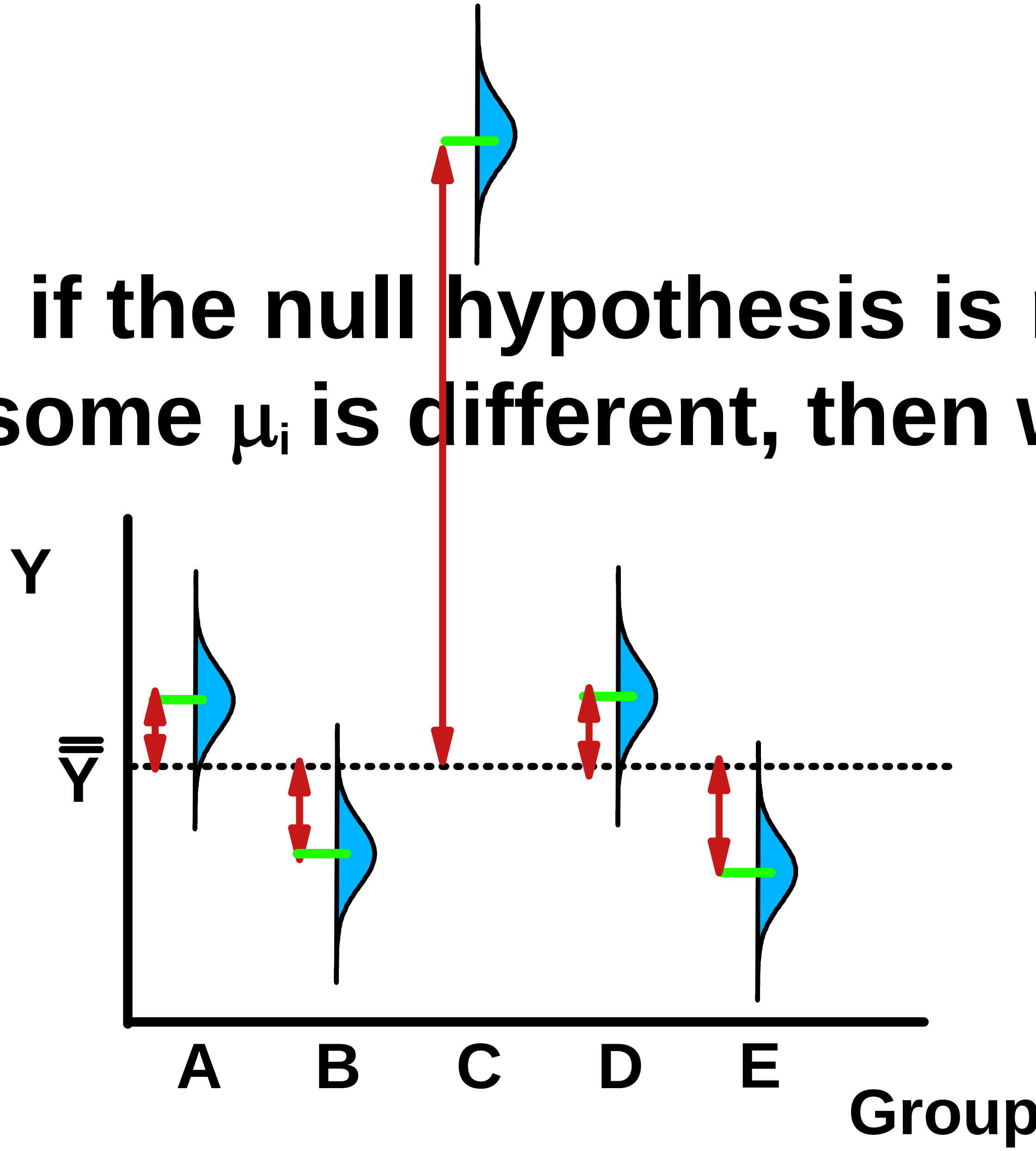
- **So if the null hypothesis is true, the mean square of the deviations within groups should be equal to the mean square of the deviations of the means multiplied by "n"!!!!**

ANOVA review (*continued*)



ANOVA review *(continued)*

- **Now, if the null hypothesis is not true, and some μ_i is different, then what?**

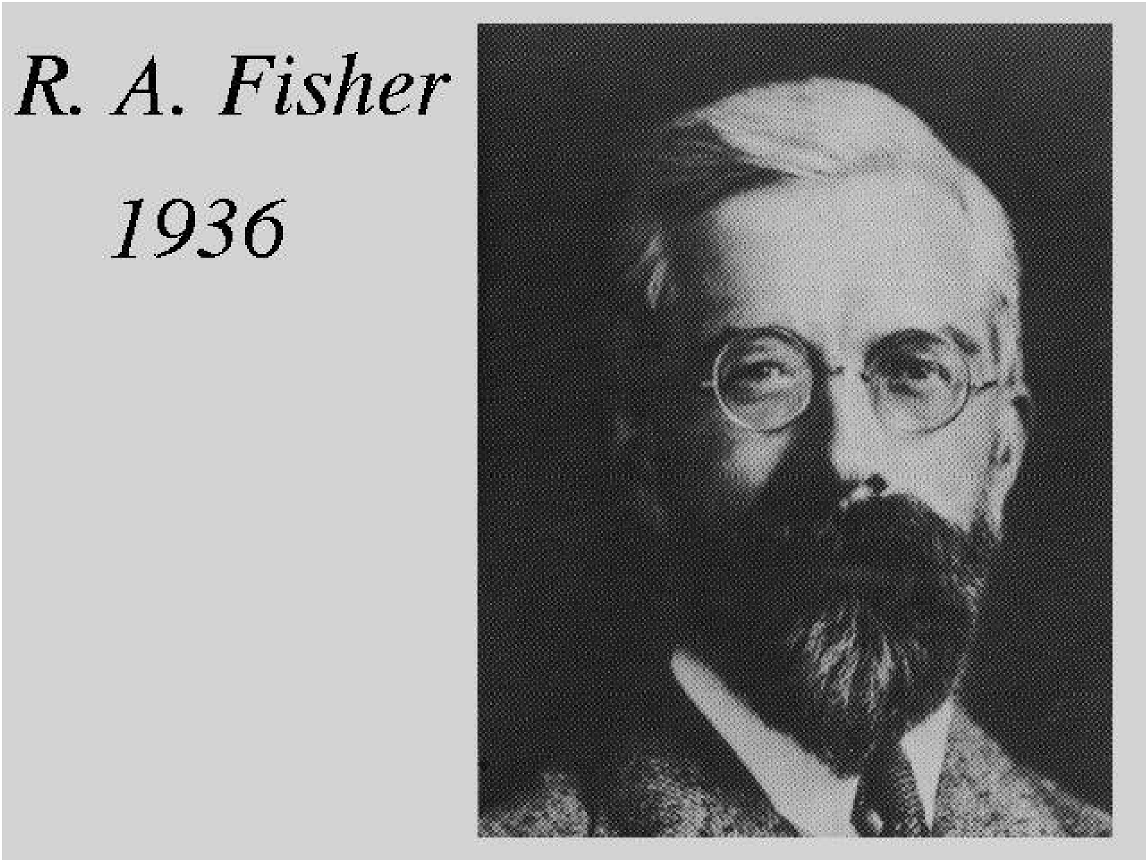


ANOVA review *(continued)*

- Then, when we calculate a mean square of deviations of the means from the overall mean, it should be larger than the previously estimated S^2_p.
- So we have two estimates of variance, S^2_p and the variance from the treatment means. If the null hypothesis is true, they should not be significantly different.

ANOVA review *(continued)*

- **If the null hypothesis is FALSE, the treatment mean square should be larger. It will therefore be a ONE TAILED TEST!**



ANOVA review *(continued)*

- We usually present this in an "Analysis of Variance" table.

Source	d.f.	Sum of Squares	Mean Square
Treatment	t-1	SSTreatment	MSTreatment
Error	t(n-1)	SSError	MSError
Total	tn-1	SSTotal	

ANOVA review *(continued)*

- **Degrees of freedom**
  - **There are tn observations total (Σn_i if unbalanced).**
  - **After the correction factor, there are tn-1 for the corrected total.**
  - **There are t-1 degrees of freedom for the t treatment levels.**
  - **Each group contributes n-1 d.f. to the pooled error term. There are t groups, so the pooled error (MSE) has t(n-1) d.f.**

ANOVA review *(continued)*

- **The SSTreatments is the SS deviations of the treatment means from the overall mean.**
- **Each deviation is denoted τ_i, and is called an treatment "effect".**

$$\text{SSTreatments} = \sum_{i=1}^{t} (\overline{Y}_i - \overline{\overline{Y}})^2 = \sum_{i=1}^{t} \tau_i^2$$

ANOVA review *(continued)*

- The model for regression is
  - $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- The effects model for a CRD is
  - $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$
    - where the treatments are i=1, 2, ... t
    - and the observations are j=1, 2, ... n, or n_i for unbalanced data
- The means model is $Y_{ij} = \mu_i + \varepsilon_{ij}$

ANOVA review *(continued)*

- **The calculations.**
- **The SSTotal is exactly the same as regression, the sum of all ΣY^2_{ij} observations (squared first).**
- **The correction factor is exactly the same too, all observations are summed, the sum is squared and divided by the number of observations, $(\Sigma Y_{ij})^2/tn$.**
- $$\text{UncorrectedSSTreatments} = \sum_{i=1}^{t}\left(\sum_{j=1}^{n} Y_{ij}\right)^2$$

ANOVA review *(continued)*

Obs	Group 1	Group 2	Group 3	Group 4
1	Y_{11}	Y_{21}	Y_{31}	Y_{41}
2	Y_{12}	Y_{22}	Y_{32}	Y_{42}
...				
n	Y_{1n}	Y_{2n}	Y_{3n}	Y_{4n}
sum	ΣY_1	ΣY_2	ΣY_3	ΣY_4
mean	$\bar{Y}_1$	$\bar{Y}_2$	$\bar{Y}_3$	$\bar{Y}_4$

ANOVA review *(continued)*

- Calculations are the same as regression for the corrected sum of squares total.
- The corrected SS treatments is the uncorrected treatments calculated from the marginals less the same correction factor used for the total.
- Error is usually calculated as the SSTotal minus the SSTreatments.

ANOVA review *(continued)*

- **We use an F test to test the equality of two means. An Analysis of Variance usually proceeds with an F test of the MSTreatment using the MSError. The test has t-1 and t(n-1) degrees of freedom.**
- **This F test will be ONE TAILED since we expect the treatment variance to be too large if the null hypothesis is not true.**

ANOVA review *(continued)*

- **The MSError estimate a variance we designate σ^2 or σ^2_ε.**
- **If the null hypothesis is true, the MSTreatments estimates the SAME VARIANCE, σ^2.**
- **However, if the null hypothesis is false the MSTreatment variance is the same σ^2 plus some amount due to the differences between treatments. This is designated $\sigma^2 + n\sigma^2_\tau$.**

ANOVA review *(continued)*

- **Since the treatment variance can be designated $\sigma^2 + n\sigma^2_\tau$, we can see that the null hypothesis can be stated as either the usual $H_0:\mu_1=\mu_2=\ldots=\mu_t$, or as $H_0:\sigma^2_\tau=0$.**
- **Which is best depends on the nature of the treatment. If the treatment levels are randomly chosen from a large number of treatment levels, then they estimate the variance of that treatment population and would be random. This would be σ^2_τ.**

ANOVA review *(continued)*

- **However, if the treatments are not chosen from a large number of treatments; if they are either all of the levels of interest or all of the levels that exist, then they are said to be FIXED. Fixed treatment levels represent a group of means that are of interest to the investigator, so $H_0{:}\mu_1=\mu_2=...=\mu_t$ is a better representation of the null hypothesis than $H_0{:}\sigma^2_\tau=0$.**

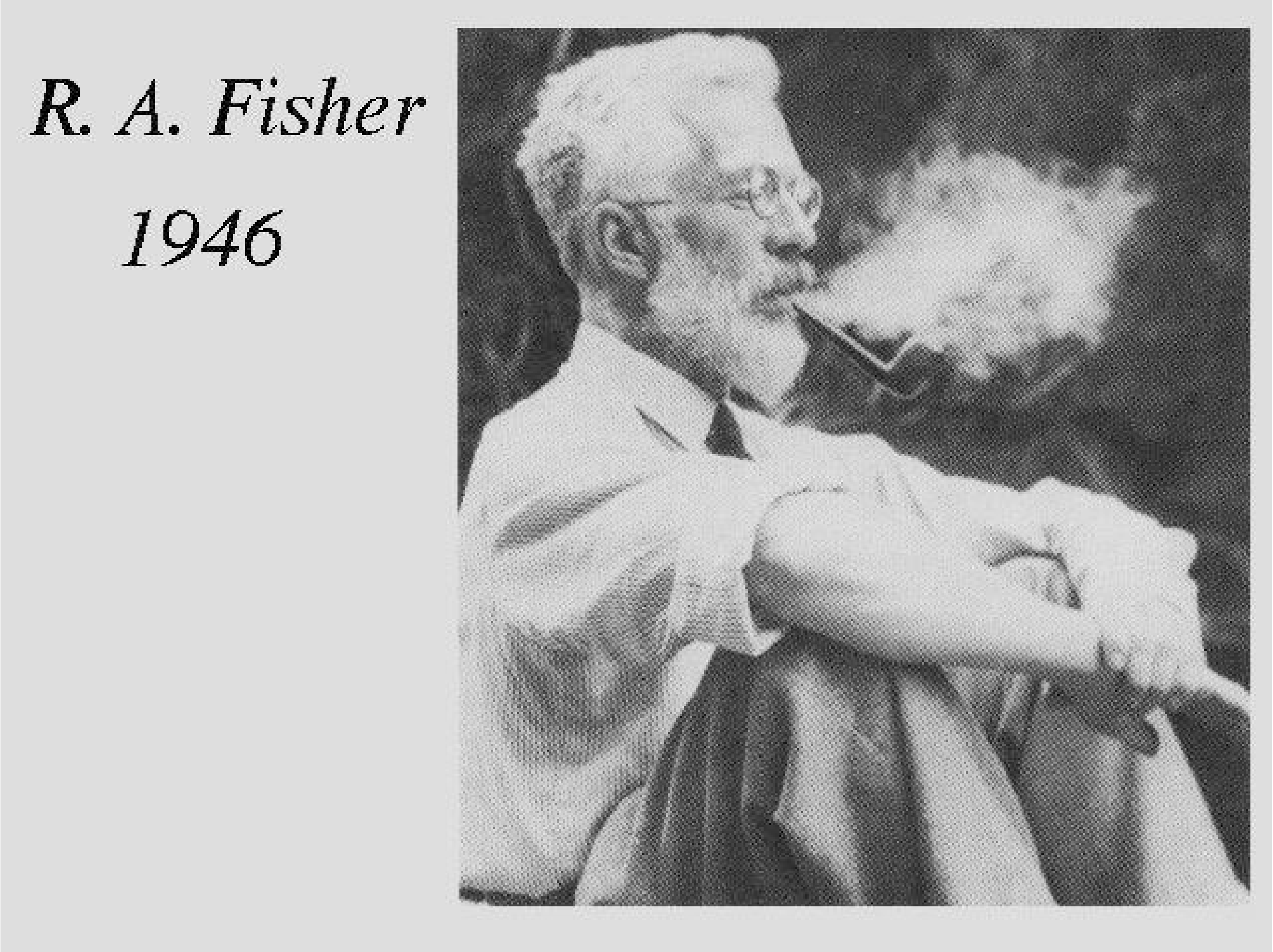
ANOVA review *(continued)*

- **For fixed treatments we still calculate a sum of squared treatment effects and divide by d.f. This is designated $n\Sigma\tau^2_i,/(n-1)$ and the F test is the same. These simply do not represent a variance.**
- **The two values estimated by the MSTreatment ($\sigma^2 + n\sigma^2_\tau$ or $\sigma^2 + n\Sigma\tau^2_i,/(n-1)$) and MSError ($\sigma^2$) are called <u>expected mean squares</u>.**

ANOVA review *(continued)*

- **One final note on the F test.**
- **Given that MSTreatments and MSError estimate these EMS (expected mean squares) we can rewrite the F test as $F=(\sigma^2 + n\sigma^2_\tau)/(\sigma^2)$.**
- **From this value we can see that it must be a one tailed test because $n\sigma^2_\tau$ cannot be negative, so the ratio is always >1. We can also see that increasing n increases power.**

Example

- **See SAS handout.**
- 

Summary?