BAN 602: Quantitative Fundamentals

Lecture 6: ANOVA and SLR

Agenda

1. Big Picture of Chapter 13: ANOVA (Analysis of Variances)
2. Big Picture of Chapter 14: SLR (Simple Linear Regression)
3. Example 1 (ANOVA)

Big Picture of Chapter 13: ANOVA

Introduction	ANOVA and Completely Randomized Design	Multiple comparison procedures	Randomized Block Design and Factorial Experiment
• Terminology: factor, treatment, single-factor experiment, response variable, experimental units, completely randomized design • Hypothesis test ○ H_0: all population means are equal • ANOVA assumptions and overview	• Between-treatments estimate of population variance (MSTR) • Within-treatments estimate of population variance (MSE) • F test statistic: MSTR/MSE (always upper-tailed test) • (one-way) ANOVA table •	• To determine where the differences among means occur • Fisher's LSD (least significant difference) ○ Hypotheses ○ T test statistic ○ LSD (MOE) • Type I Error rate	• Randomized block design • (Two-way) ANOVA

Big Picture of Chapter 14: SLR

SLR Model, LSM (least square method) and Coefficient of Determination (R-squared)	Model Assumptions and Testing for Significance	Using the Estimated Regression Equation for Estimation and Prediction	Residual Analysis
• model ○ $Y = \beta_0 + \beta_1 X + \epsilon$ • Regression equation ○ $E(Y) = \beta_0 + \beta_1 X$ • Estimated regression equation ○ $\hat{Y} = b_0 + b_1 X$ • Min SSE • SST = SSR + SSE ○ R^2 = SSR/SST • Correlation coefficient = (sign of b1)*sqrt(R^2) •	• The error term is independently, identically, and normally distributed with zero mean and standard deviation σ. • Estimate of σ: sqrt(MSE), called SE of the estimate • T test (sampling distribution of b_1) • F test (MSR/MSE) • ANOVA for SLR	• Interval estimation ○ Confidence interval (interval estimate of the mean value of Y for a given value of X) ○ Prediction interval (predict an individual value of Y for a given X) ○ Pay attention to the difference in standard error for confidence and prediction interval •	• Residual plots • Standardized residuals • Normal probability plot • outliers • influential observation

Example 1

- Chemitech developed a new filtration system for municipal water suppliers. The components for the new filtration system will be purchased from several suppliers, and Chemitech will assemble the components at its Columbia, SC plant. The industrial engineering group is responsible for determining the best assembly method for the new filtration system. After considering a variety of possible approaches, the group narrows the alternative to three: method A, method B, and method C. These methods differ in the sequence of steps used to assemble the system. Managers want to determine which assembly method can produce the greatest number of filtration systems per week.

Example 1

- Here, assembly method is the independent variable, also called **factor**. The three assembly methods that correspond to this factor are called three **treatment**. The Chemitech problem is an example of **single-factor experiment**, involving one qualitative factor (method of assembly).
- The three assembly methods or treatments define the three populations: one population is all employees who use method A, another is those who use method B, and the third is those who use method C.
- For each population, the dependent or response variable is the number of filtration systems assembled per week.
- The primary statistical objective is to determine whether the mean # of units produced per week is the same for all three populations or methods.

Example 1

- Suppose that 15 workers are randomly selected and assigned to each of the three methods equally (5 workers per method or treatment). And the following data is collected: (overall sample mean = 60; overall sample variance = 61.43)

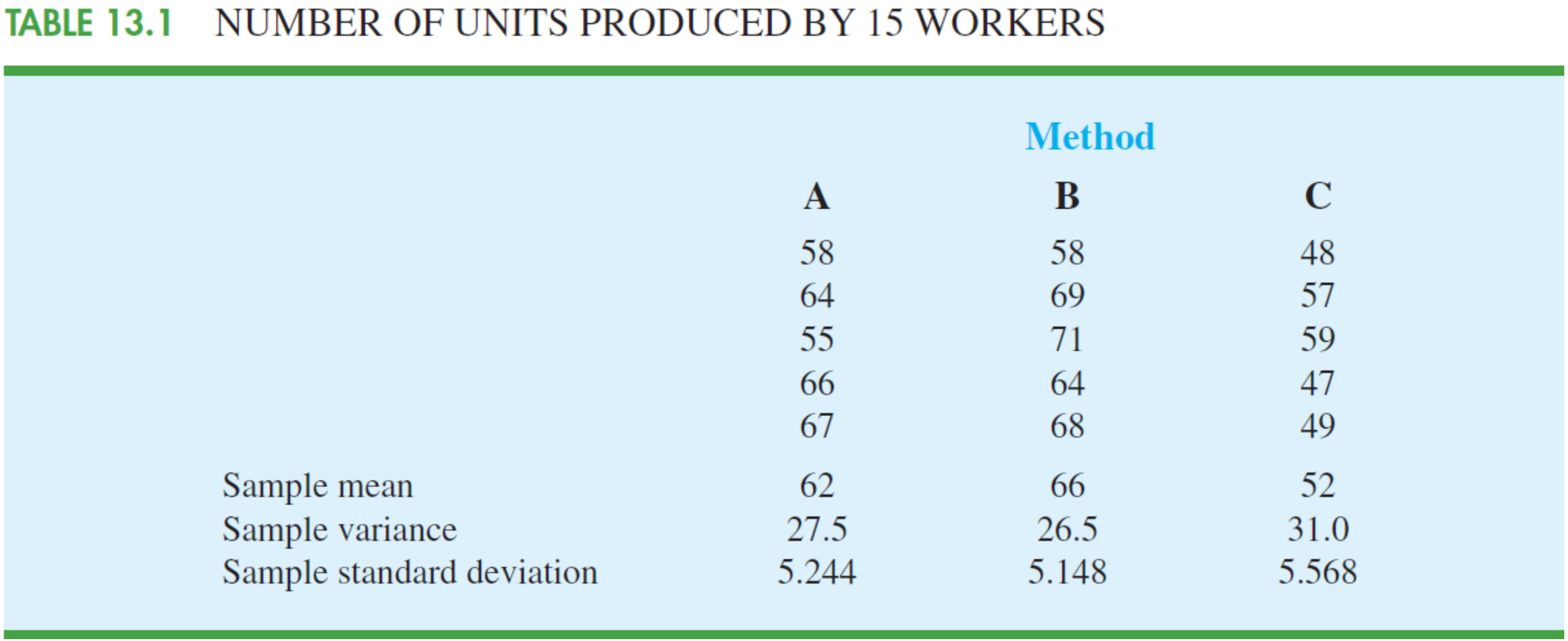


TABLE 13.1 NUMBER OF UNITS PRODUCED BY 15 WORKERS

	Method		
	A	B	C
	58	58	48
	64	69	57
	55	71	59
	66	64	47
	67	68	49
Sample mean	62	66	52
Sample variance	27.5	26.5	31.0
Sample standard deviation	5.244	5.148	5.568

Example 1

- How can we determine whether a single factor (assembly method) has an impact on the response variable (mean # of filtration systems produced)?

Example 1

- How can we determine whether a single factor (assembly method) has an impact on the response variable (mean # of filtration systems produced)?

    1. Randomly assign workers to different groups, each using a different method of assembly. A factor is essentially an independent variable. In this example, this factor/variable is categorical, having three possible values.

    2. Measure the number of units produced by each worker. Compute group/method means and variances (standard deviations). They are sample means and sample SDs.

    3. If the group means are significantly different, then the factor, assembly method, has an impact.

Example 1

- Notation
  - μ_1 = mean # of units produced per week using method A
  - μ_2 = mean # of units produced per week using method B
  - μ_3 = mean # of units produced per week using method C
- Hypotheses:
  - H_0: $\mu_1 = \mu_2 = \mu_3$
  - H_a: not all population means are equal
- ANOVA is the statistical procedure used to determine whether the observed differences in the three sample means are large enough to reject H_0.

Example 1

- How can we determine whether the population means are the same or significantly different?

    1. If the population means are the same, then the group mean (or sample mean for each treatment) should be rather close to the overall sample mean (grand mean).

Example 1: Assumptions for ANOVA

1. **For each population, the response variable is normally distributed**. (In the Chemitech example, the number of units produced per week must be normal distributed for each assembly method.)
2. **The variance of the response variable, σ^2 is the same for all of the populations**. (In the Chemitech example, the variance of the number of units produced per week must be the same for each assembly method.)
3. **The observations must be independent**. (In the Chemitech example, the number of units produced per week for each employee must be independent of the number of units produced per week for any other employee.)

Example 1: General ANOVA with Completely Randomized Design

- If the variability among the sample means is small, it supports H_0; if the variability among the sample means is large, it supports H_a.
- Hypotheses: test for the equality of k population means

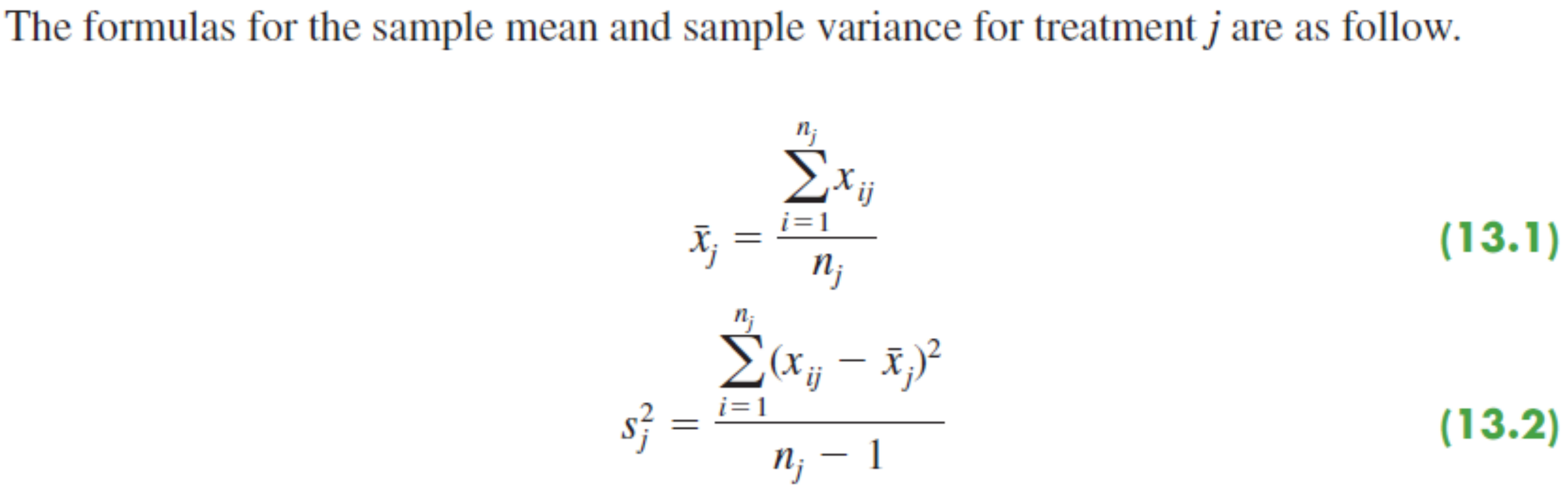
  ❑ $H_0\colon \mu_1 = \mu_2 = \cdots = \mu_k$

  ❑ H_a: not all population means are equal

Example 1: General ANOVA with Completely Randomized Design

- Notations
  - ❑ μ_j: mean of the jth population
  - ❑ X_{ij}: the value of observation i for treatment j
  - ❑ n_j: number of observations for treatment j
  - ❑ $\bar{X}_j$: sample mean for treatment j
  - ❑ s_j: sample standard deviation for treatment j

Example 1

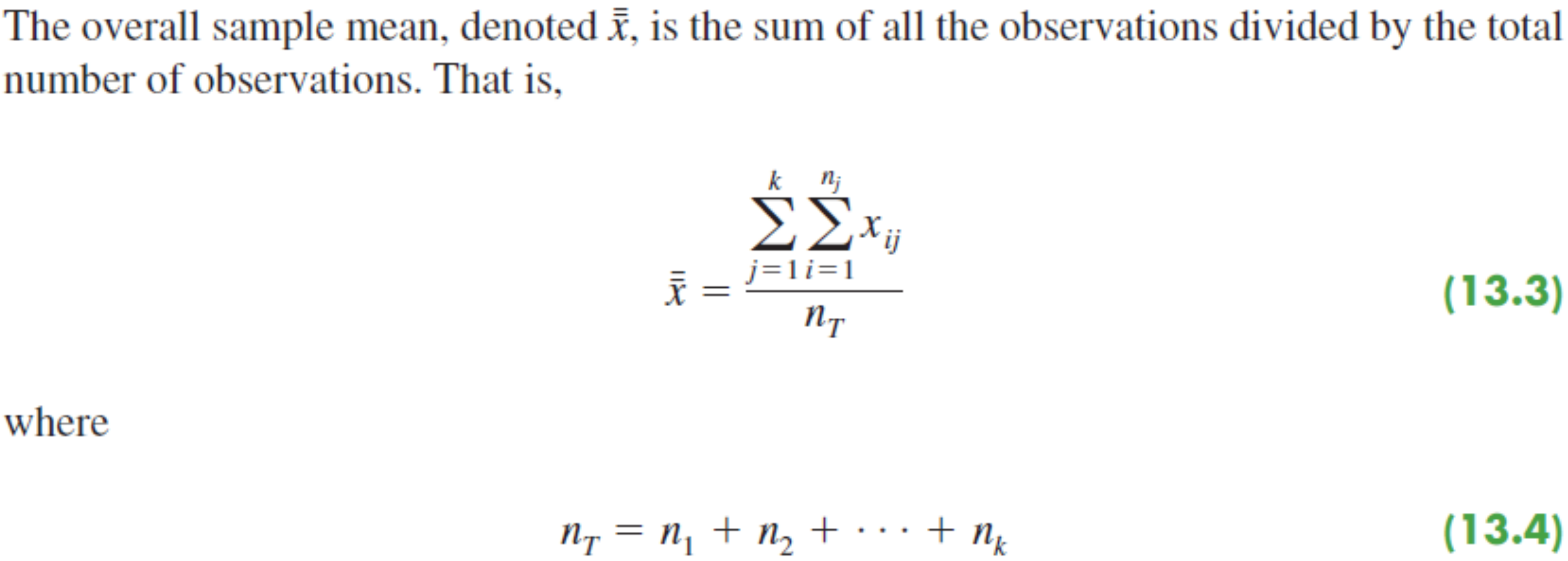


The formulas for the sample mean and sample variance for treatment *j* are as follow.

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} \tag{13.1}$$

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_j - 1} \tag{13.2}$$

Example 1



The overall sample mean, denoted $\bar{\bar{x}}$, is the sum of all the observations divided by the total number of observations. That is,

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{n_T} \tag{13.3}$$

where

$$n_T = n_1 + n_2 + \cdots + n_k \tag{13.4}$$

Example 1

- Apparently in practice, the actual number of units each worker produces will be more or less different from the overall sample mean. What are the sources of the variation/difference?

Example 1

- Apparently in practice, the actual number of units each worker produces will be more or less different from the overall sample mean. What are the sources of the variation/difference?

1. The variation may be caused by the assembly methods.

2. The variation may be caused by other factors such as age, education, gender, etc. These factors are considered as “noise”. In a perfectly designed random experiment, the impact of noises (called random errors) on the response variable will more or less cancel each other out.

3. Therefore, if the assembly method matters, than the variation caused by assembly method should be large, while the variation caused by “noise” should be small.

Example 1

- How can we measure or quantify the total variation (variance)?

Example 1

- How can we measure or quantify the total variation?

    1. Sum of squared deviation of units assembled by each worker from the overall sample mean. (SST)

    2. It is equal to (n-1)*overall sample variance.

Example 1

- How can we measure or quantify the variation caused by assembly method?

Example 1

- How can we measure or quantify the variation caused by assembly method?

    1. Replace each value with its corresponding group mean (erase individual differences or noises).
    2. Sum of squared deviation of each group mean from the overall sample mean. (SSTR)

Example 1

- How can we measure or quantify random errors or the variation caused by noises?

Example 1

- How can we measure or quantify random errors or the variation caused by noises?

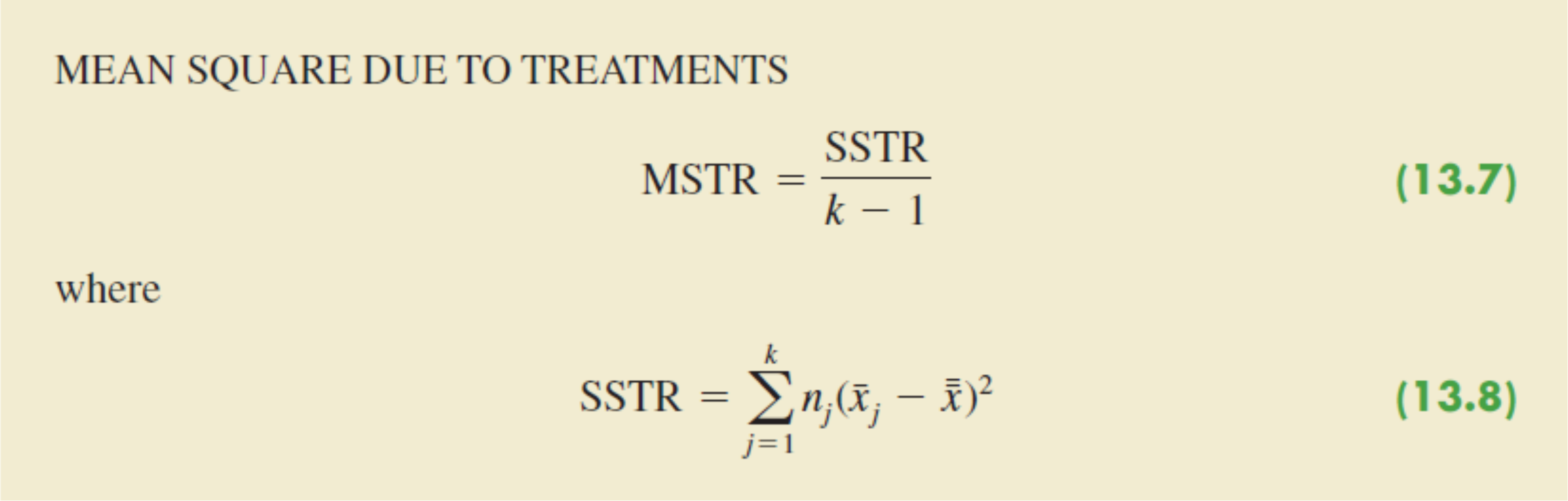
    1. For each group or treatment, compute sum of squared deviation of units assembled by each worker from the group sample mean. This is equal to (group or treatment sample size – 1)*treatment sample variance.

    2. Sum of the squared deviation across all groups or treatment. (SSE)

Example 1

- Interesting fact: SST = SSTR + SSE
- MST = SST/(overall sample size – 1)
- MSTR = SSTR/(# of treatment – 1)
- MSE = SSE/(overall sample size - # of treatment)
- F test statistic = MSTR/MSE

Example 1



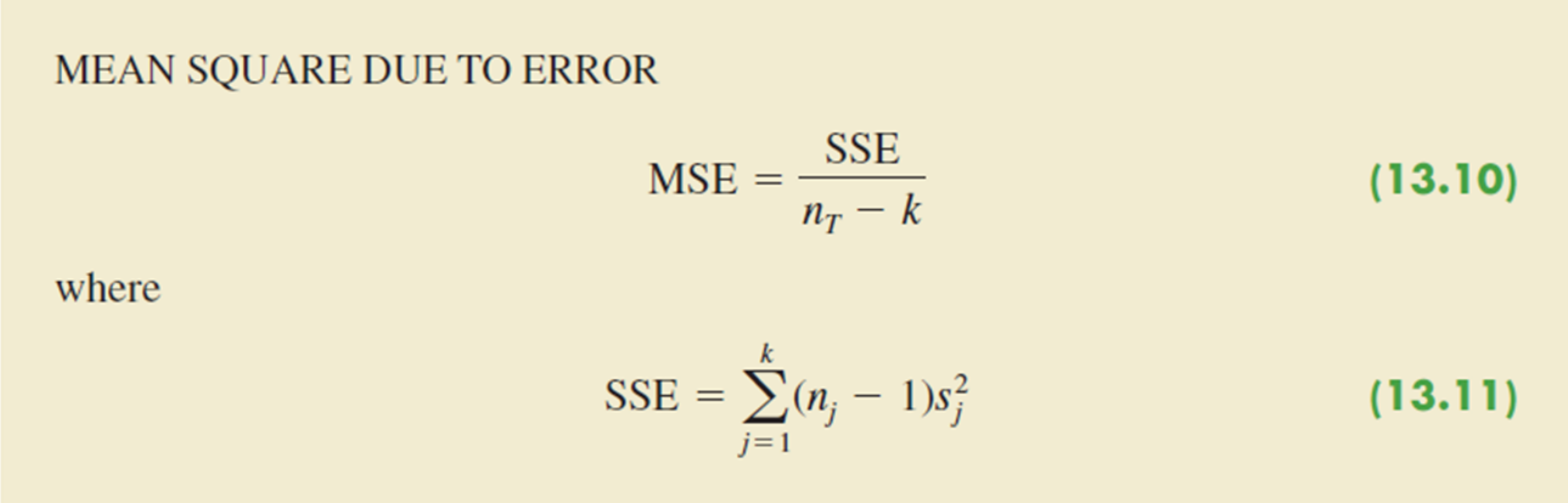
MEAN SQUARE DUE TO TREATMENTS

$$\text{MSTR} = \frac{\text{SSTR}}{k - 1} \qquad (13.7)$$

where

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 \qquad (13.8)$$

Example 1



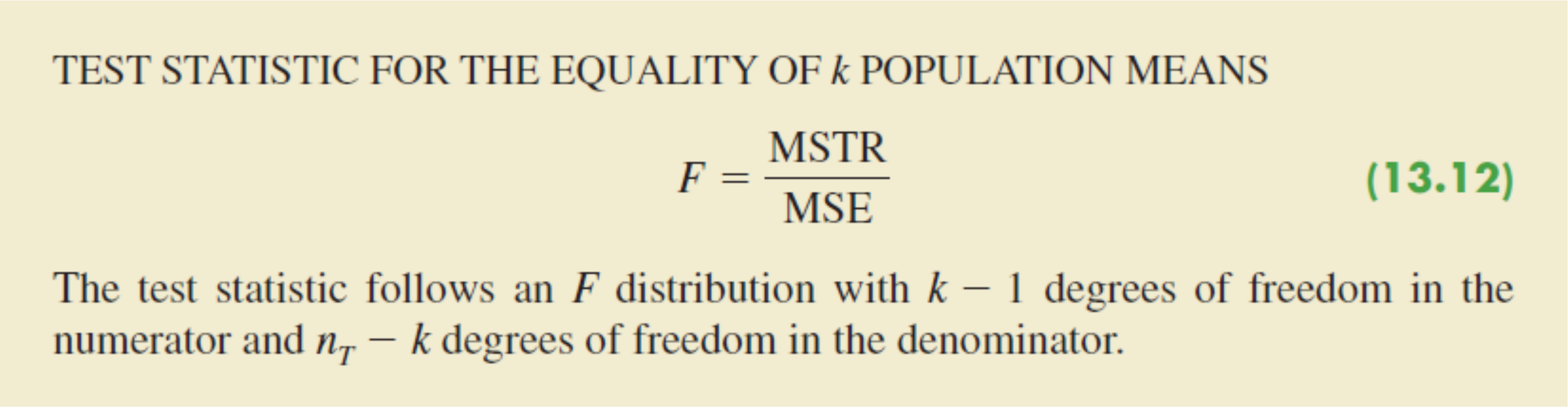
MEAN SQUARE DUE TO ERROR

$$\text{MSE} = \frac{\text{SSE}}{n_T - k} \qquad \textbf{(13.10)}$$

where

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 \qquad \textbf{(13.11)}$$

Example 1

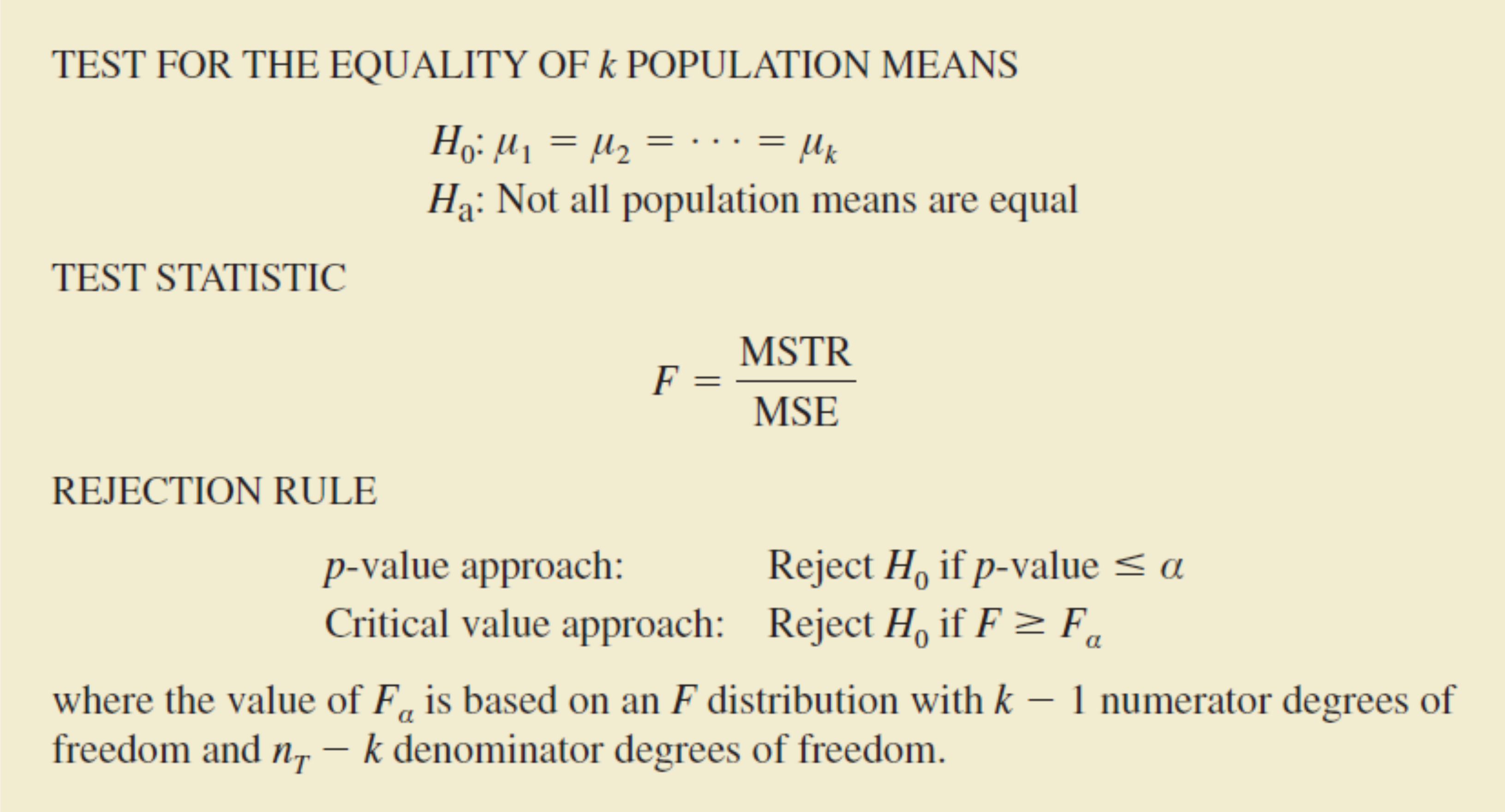


TEST STATISTIC FOR THE EQUALITY OF k POPULATION MEANS

$$F = \frac{\text{MSTR}}{\text{MSE}} \tag{13.12}$$

The test statistic follows an F distribution with $k - 1$ degrees of freedom in the numerator and $n_T - k$ degrees of freedom in the denominator.

Example 1



TEST FOR THE EQUALITY OF *k* POPULATION MEANS

$$H_0\colon \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_a\colon \text{Not all population means are equal}$$

TEST STATISTIC

$$F = \frac{\text{MSTR}}{\text{MSE}}$$

REJECTION RULE

p-value approach: Reject H_0 if p-value $\leq \alpha$

Critical value approach: Reject H_0 if $F \geq F_\alpha$

where the value of F_α is based on an F distribution with $k - 1$ numerator degrees of freedom and $n_T - k$ denominator degrees of freedom.

Example 1

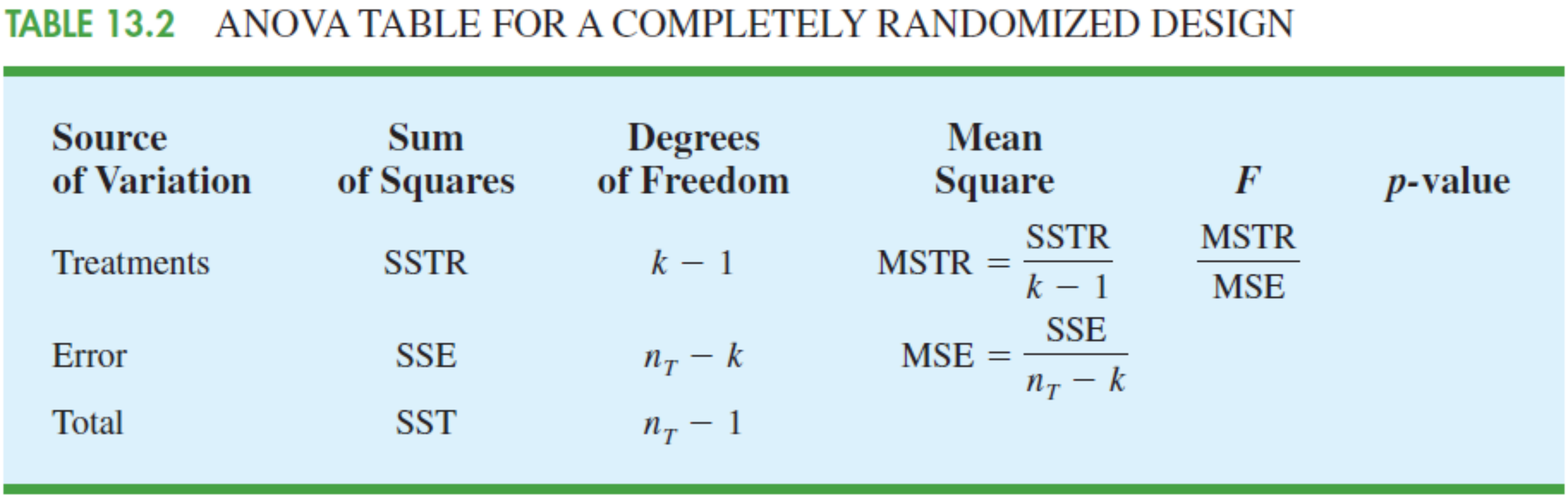


TABLE 13.2 ANOVA TABLE FOR A COMPLETELY RANDOMIZED DESIGN

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
Treatments	SSTR	$k - 1$	$\text{MSTR} = \dfrac{\text{SSTR}}{k - 1}$	$\dfrac{\text{MSTR}}{\text{MSE}}$	
Error	SSE	$n_T - k$	$\text{MSE} = \dfrac{\text{SSE}}{n_T - k}$		
Total	SST	$n_T - 1$			

ANOVA

Source of Variation	Sum of Squares (SS)	Degrees of freedom (df)	Mean Squares (MS)	F	P-value
Treatment					
Error				NA	NA
Total			NA	NA	NA

ANOVA of Example 1

Source of Variation	Sum of Squares (SS)	Degrees of freedom (df)	Mean Squares (MS)	F	P-value
Treatment	520	2	260	9.18	0.0038
Error	340	12	28.33	NA	NA
Total	860	14	NA	NA	NA

ANOVA of Example 1

- R code for computing p-value
- R code for computing critical F value (alpha = 5%)
- Acceptance/rejection region for the null hypothesis

ANOVA of Example 1

- R code for computing p-value: 1 – pf(9.18, 2, 12) = 0.38%
- R code for computing critical F value: qf(0.95, 2, 12) = 3.89
- Acceptance/rejection region for the null hypothesis
  - Acceptance: [0, 3.89)
  - Rejection: [3.89, infinity)

Quiz 6

- Multiple proportions and applications
    - understand/formulate null and alternative hypothesis
    - Test statistic and sampling distribution
    - R code for p-value and critical value of test statistic
- Multiple means
    - understand/formulate null and alternative hypothesis
    - Test statistic and sampling distribution
    - R code for p-value and critical value of test statistic
    - Complete ANOVA table with sample means and sample variances provided.