Stat 445/545: Analysis of Variance and Experimental Design

Chapter 16: Single-Factor Studies

Instructor: Yan Lu

THE UNIVERSITY OF NEW MEXICO

Example 1: A hospital research staff wished to determine the best dosage level among three dosages levels for a standard type of drug therapy to treat a medical condition.

- 30 patients with the medical problem were recruited to participate in a pilot study.
- Each patient was randomly assigned to one of the three drug dosage levels
  ——with exactly 10 patients studied in each drug dosage level group
  ——the design is balanced, because each treatment is replicated the same number of times

This is an example of balanced completely randomized design, based on a single, three-level quantitative factor.

Example 2: Want to investigate absorptive properties of four different formulations of a paper towel.

- Five sheets of paper towel were randomly selected from each of the four types, total of 20 paper towels.
- Twenty 6-ounce beakers of water were prepared, and the 20 paper towels were randomly assigned to the beakers.
  ——paper towels were fully submerged in the beaker water for 10 seconds, withdrawn, and the amount of water absorbed by each paper towel sheet was determined and recorded

This is an example of balanced completely randomized design, based on a single, four-level quantitative factor.

Example 3: Four machines in a plant were studied with respect to the diameters of ball bearings they produced.

- the purpose of the study was to determine whether substantial differences in the diameters of ball bearings existed betweent the machines.

This is an observational study, as no randomization of treatments to experimental units occured.

Relation between Regression and Analysis of Variance (ANOVA)

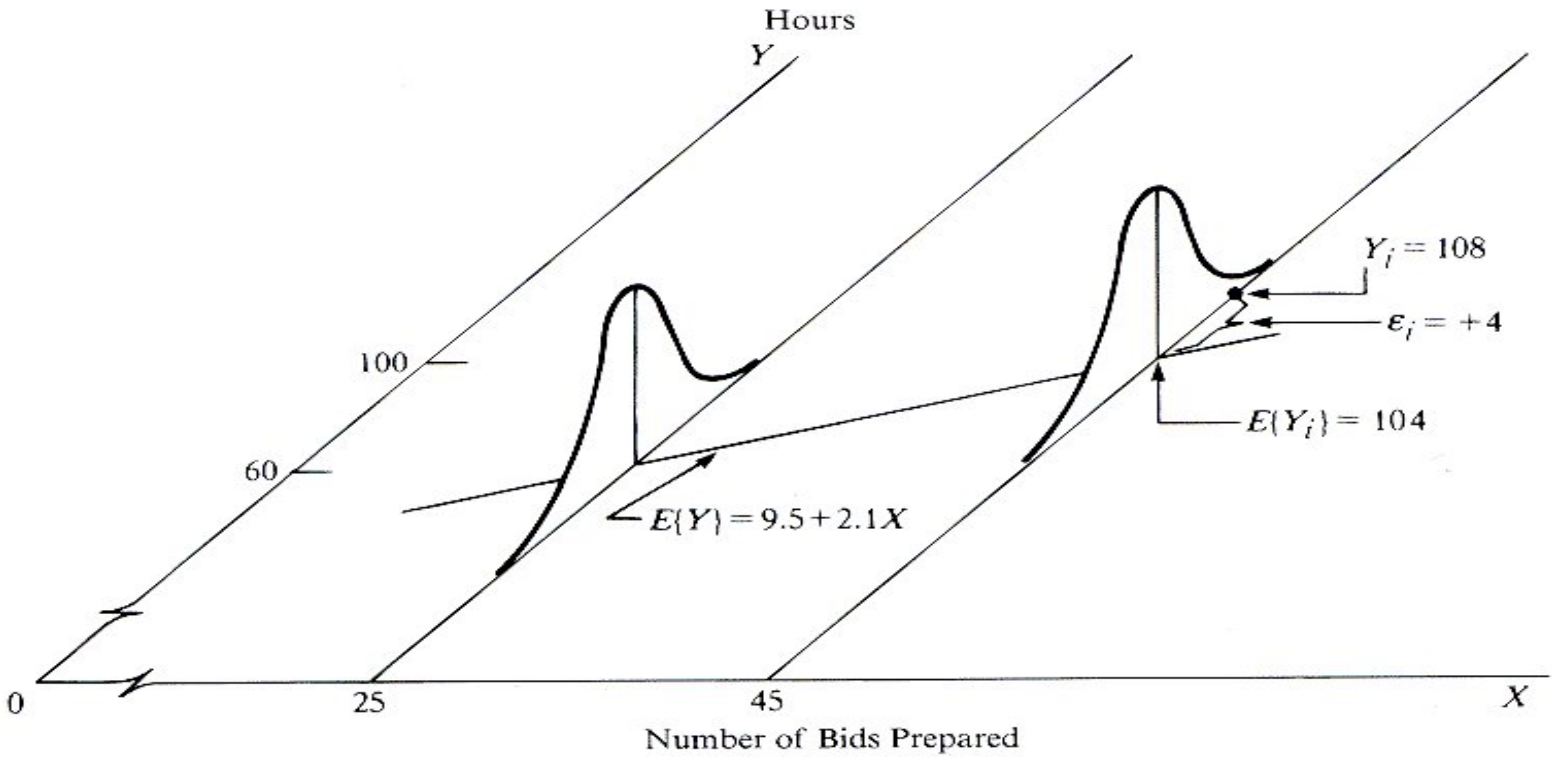
- Both Regression and ANOVA models are concerned with the statistical relation between one or more predictor variables and a response variable.
- Regression:
  ——In ordinary regression, both the predictor and response variables are quantitative.
  ——The regression function describes the nature of the statistical relation between the mean response and the levels of the predictor variable(s).
- ANOVA:
  —-The response varialbe is continuous (quantitative)
  —-The predictor variables are usually qulitative/categorical (gender, geographic location, plant shift, etc.)
  —-If the predictor variables are quantitative, no assumption is made in ANOVA model about the nature of the statistical relation between them and the response variable.

Example: a consultant for an electrical distributor is studying the relationship between the number of bids requested by sonstruction contractors for basic lighting equipment during a week and the time required to prepare the bids.

$$E(Y) = 9.5 + 2.2X$$

- X is the number of bids prepared in a week
- Y is the number of hours required to prepare the bids
- suppose $X = 45$, then $E(Y) = 9.5 + 2.2 * 45 = 104$

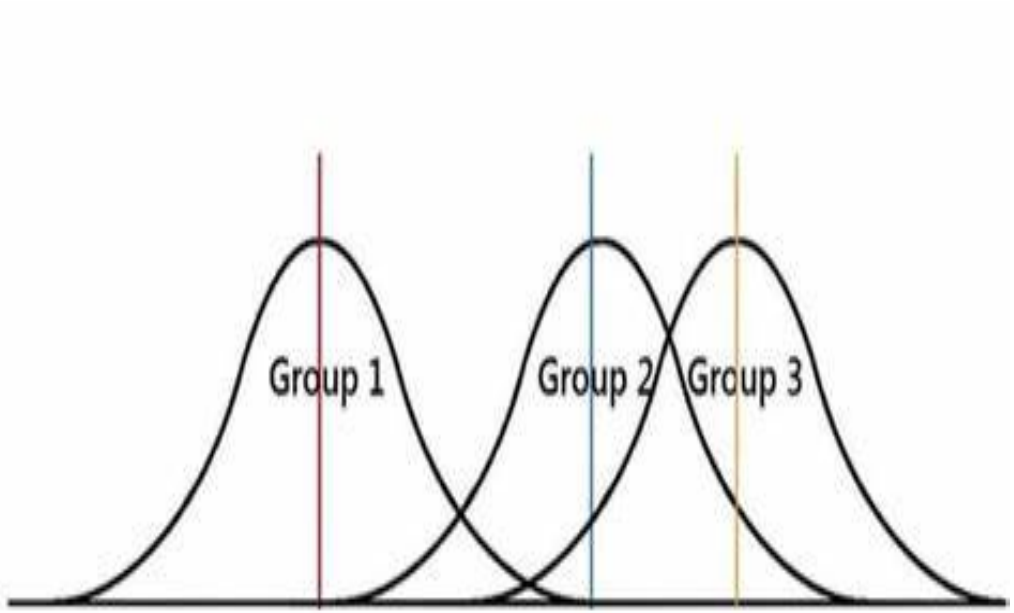
FIGURE 1.6 Illustration of Simple Linear Regression Model (1.1).



- For each level of the predictor variable, there is a probability distribution of number of hours required.
- The mean of these probability distributions fall on the regression curve, which describes the statistical relation between hours required and number of bids prepared in a week

Example: ANOVA model for a study of the effects of three different types of incentive pay systems on employee productivity.

- each type of incentive pay system corresponds to a different population, each with a probaiblity distribution of emplyee productivities (Y)
- there is no regression model representation
- want to compare the means of employee productivity by the three incentive pay system


Group 1
Group 2
Group 3

Notation:

1. Factor: an independent or predictor variable to be studied for its effect
   ——ex: dosage
2. Factor level: a particular form of that factor
   ——ex: three dosage levels
3. Single-factor analysis: only one factor involved
   —-ex: dosage example
   Multi-factor analysis: at least 3 factors involved
   —-ex: dosage (3 levels), types of medicine (2 levels), time to take medicine (2 levels, before meal or after meal)
4. Dependent or response variable: the variable that is actually measured, which is thought to depend on the value of the factor in some way.
   —-ex: pain level measured after taking medicine

5. Treatment:
   —-single-factor study: a factor level such as dosage level
   —-multi-factor study: a combination of factor levels
   X: dosage (3 levels), types of medicine (2 levels), time to take medicine (2 levels, before meal or after meal)
   total of $3 * 2 * 2 = 12$ combinations
   ex: dosage level 1 and medicine type 1 and taking medicine before meal
6. Experimental factors: the levels of the factor are assigned to experimental units randomly
7. Classification factors: a factor describes a characteristic of an experimental unit. The levels of the factor can't be assigned to the experiment.
   —-ex: formulation of paper towel

One-way ANOVA

- The one-way analysis of variance is a generalization of the two sample t-test to $r > 2$ groups.
  —— Assume that the populations of interest have the following (unknown) population means and standard deviations:

	population 1	population 2	$\cdots$	population r
mean	μ_1	μ_2	$\cdots$	μ_r
std dev	σ_1	σ_2	$\cdots$	σ_r

- A usual interest in ANOVA is whether $\mu_1 = \mu_2 = \cdots = \mu_r$. If not, then we wish to know which means differ, and by how much.

Data Structure

- Select samples from each of the r populations
- Let Y_{ij} denote the j^{th} observation in the i^{th} level/group, $i = 1, 2, \cdots, r$ and $j = 1, 2, \cdots n_i$

	sample/level 1	sample/level 2	$\cdots$	sample/level r
	$Y_{11}, \cdots, Y_{1n_1}$	$Y_{21}, \cdots, Y_{2n_2}$	$\cdots$	$Y_{r1}, \cdots, Y_{rn_r}$
size	n_1	n_2	$\cdots$	n_r
mean	$\bar{Y}_{1\cdot}$	$\bar{Y}_{2\cdot}$	$\cdots$	$\bar{Y}_{r\cdot}$
SE	S_1	S_2	$\cdots$	S_r

where $\bar{Y}_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}/n_i$.

- total sample size $n_T = n_1 + n_2 + \cdots + n_r$

- let $\bar{Y}_{\cdot\cdot}$ be the average response over all samples, that is

$$\bar{Y}_{\cdot\cdot} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{ij}}{n} = \frac{\sum_{i=1}^{r} n_i \bar{Y}_{i\cdot}}{n}.$$

Note that $\bar{Y}_{\cdot\cdot}$ is *not* the average of the sample means, unless the sample sizes n_i are equal.

One way ANOVA model—-Cell means model

$$Y_{ij} = \mu_i + \epsilon_{ij} \text{ for } i = 1, 2, \cdots, r; j = 1, 2, \cdots, n_i \qquad (1)$$

where

- Y_{ij} is the value of the response variable in the jth trial for the ith factor level/sample/group/treatment
- μ_i's are means of response values for level i, they are parmeters to be estimated
- ϵ_{ij} are independent $N(0, \sigma^2)$, $i = 1, \cdots, r; j = 1, \cdots n_i$

——

$$E(Y_{ij}) = \mu_i, V(Y_{ij}) = \sigma^2, Y_{ij} \text{ are independent } N(\mu_i, \sigma^2)$$

——Independence assumption is usually considered to be satistified by completely randomized design.

Least square estimators

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

Consider the deviation of Y_{ij} from its expected value $[Y_{ij} - \mu_i]$

- Measure:

$$\begin{aligned} Q &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \mu_i)^2 \\ &= \sum_{j=1}^{n_1}(Y_{1j} - \mu_1)^2 + \sum_{j=1}^{n_2}(Y_{2j} - \mu_2)^2 + \cdots + \sum_{j=1}^{n_r}(Y_{rj} - \mu_r)^2 \\ &= \sum_{i=1}^{r} Q_i \end{aligned}$$

- Objective: to find estimate of μ_i, for which Q is minimum

- Differentiating with respect to μ_i, we obtain

$$\frac{dQ_i}{d\mu_i} = \sum_j -2(Y_{ij} - \mu_i)$$

Set $\sum_{j=1}^{n_i} -2(Y_{ij} - \mu_i) = 0$, so that $\sum_{j=1}^{n_i} Y_{ij} = n_i\hat{\mu}_i$, therefore

- $\hat{\mu}_i = \bar{Y}_{i\cdot}$.
- The same estimators are obtained by the method of maximum likelihood. The likelihood function is

$$L(\mu_1, \cdots, \mu_r, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_i\sum_j (Y_{ij} - \mu_i)^2\right]$$

Maximizing this likelihood function with respect to the parameters μ_i is equivalent to minimizing the sum $\sum_i \sum_j (Y_{ij} - \mu_i)^2$ in the exponent, which is the least squares criterion.

Residuals

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

with LS estimators $\hat{\mu}_i = \bar{Y}_{i\cdot}$.

- Predicted (fitted or mean) value of Y_{ij} is:

$$\hat{Y}_{ij} = \bar{Y}_{i\cdot}$$

—the fitted value $\hat{Y}_{ij}$ is not the same as Y_{ij}
—Y_{ij} is the observed value, and $\hat{Y}_{ij}$ is the predicted value

- Residual $e_{ij} = Y_{ij} - \hat{Y}_{ij}$: vertical deviation between Y_{ij} and the estimated μ_i
- Error term $\epsilon_{ij} = Y_{ij} - \mu_i$: vertical deviation between Y_{ij} and the true group mean μ_i
- Residual e_{ij} is a prediction of ϵ_{ij}
—- $e_{ij} \neq \epsilon_{ij}$
A normal scores plot or histogram of the residuals should resemble a sample from a population.

- $$\sum_{j=1}^{n_i} e_{ij} = \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.}) = 0$$

- $$\begin{aligned} Y_{ij} &= \bar{Y}_{i.} + e_{ij} \\ &= \bar{Y}_{..} + (\bar{Y}_{i.} - \bar{Y}_{..}) + e_{ij} \end{aligned}$$

- Cell mean model can also be written as

$$Y_{ij} = \mu_. + \tau_i + \epsilon_{ij},$$

where μ is the overall mean, τ_i represents the treatment effect, and ϵ_{ij} are the errors.

Example: The Kenton Food Company wished to test 4 different package designs for a new breakfast cereal.

- 20 stores, with approximately equal sales volumes, were selected as the experimental units
  —-comparable in location and sales volume.
  —-other relevant conditions, that could affect sales, such as price, amount and location of shelf space and special promotional effors, were kept the same for all the stores in the experiment.
- Each store was randomly assigned one of the package designs
  —-each package design is assigned to 5 stores
- Missing data: a fire occured in one store during the study period, so it is dropped from the study.
- Response: sales in number of cases were observed and recorded for the study period.

```
sales<-read.table(file="~/Desktop/jenn/teaching/
stat445545/data/CH16TA01.txt",
    col.names=c("salevolume","design","obs"))
```

```
> #### Example: Comparison of sales
> sales<-read.table(file="~/Desktop/jenn/teaching/stat4455
+     col.names=c("salevolume","design","obs"))
> sales
```

	salevolume	design	obs
1	11	1	1
2	17	1	2
3	16	1	3
4	14	1	4
5	15	1	5
6	12	2	1
7	10	2	2
8	15	2	3

```
9            19       2    4
10           11       2    5
11           23       3    1
12           20       3    2
13           18       3    3
14           17       3    4
15           27       4    1
16           33       4    2
17           22       4    3
18           26       4    4
19           28       4    5
```

```
> ##{Numerical summaries}
> #Calculate the mean, sd, n, and se for the four designs
> #The plyr package is an advanced way to apply a function
> #to subsets of data, splitting, applying and combining
> #data"
>
>
```

```
>  library(plyr)
>  # ddply "dd" means the input and output
> #are both data.frames
>  sales.summary <-ddply(sales,
+                       "design",
+                       function(X) {
+              data.frame( m = mean(X$salevolume),
+                          s = sd(X$salevolume),
+                          n = length(X$salevolume)
+                                    )})
>  sales.summary
```

```
  design    m        s n
1      1 14.6 2.302173 5
2      2 13.4 3.646917 5
3      3 19.5 2.645751 4
4      4 27.2 3.962323 5
```

- The mean sales per store with package design 1 are estimated to be 14.6 cases for the population of stores under study
- The fitted value for each of the observations for package design 1 is $\hat{Y}_{1j} = \bar{Y}_{1.} = 14.6$
- Residual of the first observation from design 1 is $Y_{11} - \hat{Y}_{11} = 11 - 14.6 = -3.6$
  —-all the residuals are listed in the following table

	Store (j)					
Design (i)	1	2	3	4	5	Total
1	-3.6	2.4	1.4	-0.6	0.4	0
2	-1.4	-3.4	1.6	5.6	-2.4	0
3	3.5	0.5	-1.5	-2.5	NA	0
4	-0.2	5.8	-5.2	-1.2	0.8	0
All design						0

Decompsition of Total Sum of Squares

- Given a set of measurements (Y_{ij}'s) and no information about the level associated with each Y, the Total Sum of Squares is

$$SSTO = \sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2$$

- If we have information about the factor levels and that the levels have different means, we would use the deviations $Y_{ij} - \bar{Y}_{i.}$ to assess the variation within the levels

$$Y_{ij} - \bar{Y}_{..} = \bar{Y}_{i.} - \bar{Y}_{..} + Y_{ij} - \bar{Y}_{i.}$$

  - $\bar{Y}_{i.} - \bar{Y}_{..}$: deviation of estimated factor level mean around overall mean
  - $Y_{ij} - \bar{Y}_{i.}$: deviation around estimated factor level mean

- square both sides and take summation

$$\sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2 = \sum_i n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2$$

Sum of Squares (SS)

- **SSTR**: standing for treatment sum of squares . A measure of the extent of differences between the estimated factor level means, based on the deviations of the estimated factor level means $\bar{Y}_{i.}$ around the overall mean $\bar{Y}_{..}$. The more the estimated factor level means differ, the larger will be SSTR.

$$SSTR = \sum_i n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

- **SSE**: standing for error sum of squares, a measure of the random variation of the observations around the respective estimated factor level means. The more the observations for each factor level differ among themselves, the larger will be SSE.

$$SSE = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2 = \sum_i \sum_j e_{ij}^2$$

-

$$SSTO = SSTR + SSE$$

Degrees of Freedom (df)

- The df(SSTR) is the number of groups minus one, $r-1$.
- The df(SSE) is the total number of observations minus the number of groups: $(n_1-1)+(n_2-1)+\cdots+(n_r-1)=n-r$.
- These two df add to give df(SSTO) $=(r-1)+(n-r)=n-1$.

The Mean Square for each source of variation is the corresponding SS divided by its *df*.

- $$MSTO = SSTO/(n-1)$$
- $$MSTR = SSTR/(r-1)$$
- $$MSE = SSE/(n-r)$$

ANOVA Table

Source of variation	*df*	SS	MS	E(MS)
Between treatment	$r-1$	$SSTR = \sum_i n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$MSTR = \frac{SSTR}{r-1}$	$\sigma^2 + \frac{\sum n_i(\mu_i - \mu_\cdot)^2}{r-1}$
Within treatment (Error)	$n-r$	$SSE = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$	$MSE = \frac{SSE}{n-r}$	σ^2
Total	$n-1$	$SSTO = \sum_{ij}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$		

where $\mu_\cdot = \frac{\sum_i n_i \mu_i}{n}$

The MSE is identical to the **pooled variance estimator** of σ^2 in a two-sample problem when $r = 2$.

$$\begin{aligned} & \sum_{i=1}^{2}\sum_{j=1}^{n_j}(Y_{ij} - \bar{Y}_{i.})^2/(n-2) \\ = & \frac{\frac{\sum_{j=1}^{n_1}(Y_{1j} - \bar{Y}_{1.})^2}{n_1 - 1}(n_1 - 1) + \frac{\sum_{j=1}^{n_2}(Y_{2j} - \bar{Y}_{2.})^2}{n_2 - 1}(n_2 - 1)}{n_1 + n_2 - 2} \\ = & \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \end{aligned}$$

In general,

$$\text{MSE} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \cdots + (n_r - 1)S_r^2}{n - r} = S^2_{\text{pooled}}$$

$$\text{MSE} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \cdots + (n_r - 1)S_r^2}{n - r} = S^2_{\text{pooled}}$$

is a weighted average of the sample variances.

- The MSE is known as the pooled estimator of variance, and estimates the assumed common population variance.
- If all the sample sizes are equal, the MSE is the average sample variance.
- $$E(S_i^2) = \sigma^2$$

  So that

  $$E(MSE) = \frac{1}{n - r}(n - r)\sigma^2 = \sigma^2$$

Review two models:
For $i = 1, 2, \cdots, r$ and $j = n_1, n_2, \cdots, n_r$

Cell mean model

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

Factor effect model

$$Y_{ij} = \mu_. + \tau_i + \epsilon_{ij}$$

Test of equivalence of the means

Hypothesis:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$ versus $H_\alpha :$ at least two of them are not equal

or

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_r$$

Test statistic

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

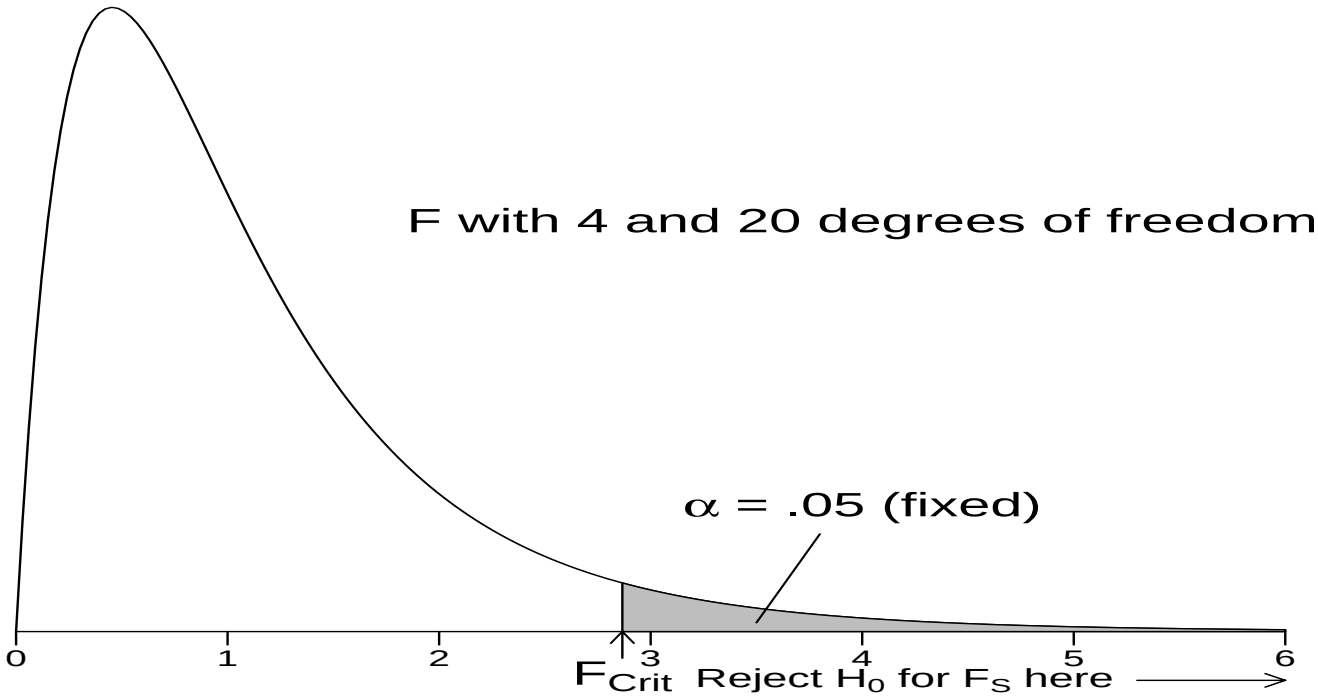
Reject H_0, if

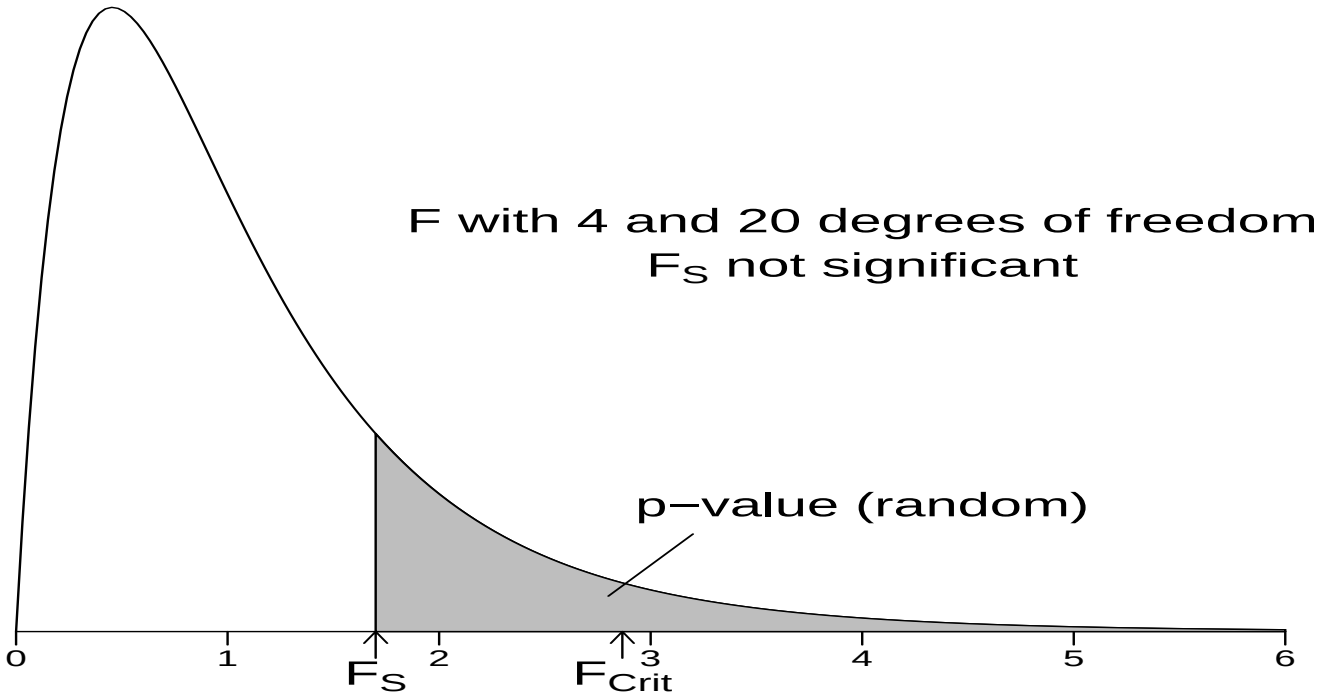
$$F^* > F(1-\alpha; r-1, n-r)$$

Do not reject H_0, if

$$F^* \leq F(1-\alpha; r-1, n-r)$$

where $F(1-\alpha; r-1, n-r)$ is the upper-α percentile from an $F(r-1, n-r)$ distribution with numerator degrees of freedom $r-1$ and denominator degrees of freedom $n-r$


F with 4 and 20 degrees of freedom
α = .05 (fixed)
0
1
2
3
4
5
6
F_Crit
Reject H0 for FS here


F with 4 and 20 degrees of freedom
F_S not significant
p−value (random)
0
1
F_S
2
F_{Crit}
3
4
5
6

Comments:

- $$SSTO = SSTR + SSE$$

  If $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$ is true, then
  ——$\bar{Y}_{i.} - \bar{Y}_{..}$ is small, i.e., SSTR is small.
  ——SSE is close to SSTO
  ——F^* is small
  If $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$ is not true, then
  ——SSTR is large
  ——SSE is small
  —- F^* is large, reject H_0.

- Large values of F^* indicate large variability among the sample means $\bar{Y}_{1.}, \bar{Y}_{2.}, \ldots, \bar{Y}_{r.}$ relative to the spread of the data within samples. That is, large values of F^* suggest that H_0 is false.

- $$E(MSE) = \sigma^2$$

$$E(MSTR) = \sigma^2 + \frac{\sum_i n_i(\mu_i - \mu_.)^2}{r-1}$$

If H_0 is true, $\mu_1 = \mu_2 = \cdots = \mu_r = \mu_.$

$$E(MSTR) = \sigma^2 = E(MSE)$$

$$F^* \approx 1$$

If H_0 is not true,

$$E(MSTR) > \sigma^2 = E(MSE)$$

$$F^* = \frac{MSTR}{MSE} > 1$$

A rough way to reject H_0.

- The p-value for the test is the area under the F-probability curve to the right of F^*.

Randomization Test

Example: two treatments (t_1, t_2), 4 experimental units (a, b, c, d), results are as follows

Treatment 1	Treatment 2
Y_{1j}	Y_{2j}
3	8
7	10

want to test if there are treatment effect or not.

$\bar{Y}_{1.} = (3+7)/2 = 5$, $\bar{Y}_{2.} = (8+10)/2 = 9$, $\bar{Y}_{..} = (3+7+8+10)/4 = 7$

$$\begin{aligned} SSTR &= \sum_{i=1}^{2} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2 \\ &= 2(5-7)^2 + 2(9-7)^2 = 16 \end{aligned}$$

$$MSTR = 16/(2-1) = 16$$

$$
\begin{aligned}
SSE &= \sum_{i=1}^{2}\sum_{j=1}^{2}(Y_{ij} - \bar{Y}_{i.})^2 \\
&= (3-5)^2 + (7-5)^2 + (8-9)^2 + (10-9)^2 = 10
\end{aligned}
$$

$$MSE = SSE/(n-r) = 10/(4-2) = 5$$

$$F^* = MSTR/MSE = 16/5 = 3.2$$

$$P(F(1,2) \geq 3.2) = 0.22$$

We fail to reject H_0 that there is no treatment effect between t1 and t2 assuming ANOVA model is applicable.

Now want to test if there are treatment effect or not without assuming the distribution of error terms (sample size of 4 is not convincing to assume a distribution).
Randomization can provide the basis for making inferences without requiring assumptions about the distribution of the error terms ϵ.

Step 1: Ranodmly assign treatment t_1 and t_2 to the four experimental units, which has a total of $4!/2!2! = 6$ assignments.
——If there is no treatment effect, the response Y_{ij} could with equal likelihood have been observed for any of the treatments.

Treatment 1 Y_{1j}	Treatment 2 Y_{2j}
3	8
7	10

	a	b	c	d			obsn		F^*	Probability
1	t1	t1	t2	t2	3	7	8	10	3.20	1/6
2	t1	t2	t1	t2	3	8	7	10	1.06	1/6
3	t1	t2	t2	t1	3	10	7	8	0.08	1/6
4	t2	t1	t2	t1	7	8	3	10	0.08	1/6
5	t2	t1	t1	t2	7	10	3	8	1.06	1/6
6	t2	t2	t1	t1	8	10	3	7	3.20	1/6

The last two columns give the randomization distribution of test statistic F^* under H_0.

From the original data set, we have $F^* = 3.20$,

$$P(F^* \geq 3.20) = 2/6 = 0.33$$

using the randomization distribution.
This P-value is somewhat different than the usual (normal theory) P-value

$$P(F(1,2) \geq 3.2) = 0.22$$

We fail to reject H_0 that there is no treatment effect between t1 and t2.

Regression approach to signle-factor analysis of variance

Factor effects model with unweighted mean

$$Y_{ij} = \mu_. + \tau_i + \epsilon_{ij} \text{ for } i = 1, 2, \cdots, r; j = 1, 2, \cdots n_i$$

where $n_1 = n_2 = \cdots n_r$,

$$\sum_{i=1}^{r} \tau_i = \sum_i (\mu_i - \mu_.) = \sum_i \mu_i - r\mu_. = 0 \qquad (2)$$

Recall that $\mu_. = \sum_{i=1}^{r} n_i \mu_i / n = \sum_{i=1}^{r} \mu_i / r$ when $n_1 = n_2 = \cdots = n_r$
Because of the restriction in (2),

$$\tau_r = -\tau_1 - \tau_2 - \cdots - \tau_{r-1},$$

we shall use only the parameters $\mu_., \tau_1, \cdots, \tau_{r-1}$ for the linear model.

An simple example for illustration:
Consider a single-factor study with $r = 3$ factor levels and with $n_1 = n_2 = n_3 = 2$. Let

- $\mathbf{X}$ be the design matrix
- β be the vector of parameters
- ε be the error vector
- $\mathbf{Y}$ be the response vector

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix}, \beta = \begin{bmatrix} \mu_. \\ \tau_1 \\ \tau_2 \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{bmatrix}$$

We have

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

$$E(\mathbf{Y}) = \begin{bmatrix} E[Y_{11}] \\ E[Y_{12}] \\ E[Y_{21}] \\ E[Y_{22}] \\ E[Y_{31}] \\ E[Y_{32}] \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu_. \\ \tau_1 \\ \tau_2 \end{bmatrix} = \begin{bmatrix} \mu_. + \tau_1 \\ \mu_. + \tau_1 \\ \mu_. + \tau_2 \\ \mu_. + \tau_2 \\ \mu_. - \tau_1 - \tau_2 \\ \mu_. - \tau_1 - \tau_2 \end{bmatrix}$$

$$\tau_3 = -\tau_1 - \tau_2, E(Y_{ij}) = \mu_. + \tau_i$$

We need to define indicator variables that take on values 0, 1 or -1.

Let X_{ij1} denote the value of indicator variable X_1 for the jth case from the ith factor level; X_{ij2} denote the value of indicator variable X_2 for the jth case from the ith factor level, and so on. The multiple regression model then is as follows:

$$Y_{ij} = \mu_. + \tau_1 X_{ij1} + \tau_2 X_{ij2} + \cdots \tau_{r-1} X_{ij,r-1} + \varepsilon_{ij}$$

where

$$X_{ij1} = \begin{cases} 1 & \text{if case is from factor level 1} \\ -1 & \text{if case is from factor level r} \\ 0 & \text{otherwise} \end{cases}$$

$$\vdots$$

$$X_{ij,r-1} = \begin{cases} 1 & \text{if case is from factor level r-1} \\ -1 & \text{if case is from factor level r} \\ 0 & \text{otherwise} \end{cases}$$

The intercept term is $\mu_.$, and the regression coefficients are $\tau_1, \tau_2, \cdots, \tau_{r-1}$.

- The least square estimator of $\mu_.$ is the average of the cell smaple means:

$$\hat{\mu}_. = \frac{\sum_{i=1}^{r} \bar{Y}_{i.}}{r}$$

——this quantity is generally not the same as the overall mean $\bar{Y}_{..}$ unless the cell smaple sizes are equal.

- The least square estimators of the ith factor effect is

$$\hat{\tau}_i = \bar{Y}_{i.} - \hat{\mu}_.$$

- To test the equality of the treatment means μ_i by means of the regression approach,

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_{r-1} = 0 \text{ versus not all } \tau_i \text{ equal zero}$$

Reduced model is therefore

$$Y_{ij} = \mu_. + \varepsilon_{ij}$$

The general linear test for whether there is a regression relation is

$$\begin{aligned} F^* &= \frac{\frac{SSE(R) - SSE(F)}{dfE(R) - dfE(F)}}{\frac{SSE(F)}{dfE(F)}} = \frac{\frac{SSTO - SSE}{n-1-(n-r)}}{MSE} \\ &= \frac{SSTR/(r-1)}{MSE} = \frac{MSTR}{MSE} \end{aligned}$$

Reject H_0, when $F^* > F(1-\alpha; r-1, n-r)$