Probability & Statistics for IS

Chapter 3: Univariate and Multivariate ANOVA

Lecturers

Dr. Ghandi Manasra
Dr. Monjed H. Samuh

Palestine Polytechnic University
(ghandi@ppu.edu, monjedsamuh@ppu.edu)

Term 191

Table of Contents

Learning Objectives

After studying this chapter, the student will:

- be able to use R to model basic experimental designs.
- fit and interpret ANOVA type models.
- evaluate model assumptions.

One-Way ANOVA: Introduction

- ANOVA stands for "**ANalysis Of VAriance**".
- The term ANOVA is a little misleading. Although the name of the technique refers to variances, the main goal of ANOVA is to investigate **differences in means**.
- One-way ANOVA is an extension of the independent *t*-test.
- The One-way ANOVA can test the equality of several population means. That is:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{"No treatment effect"}$$

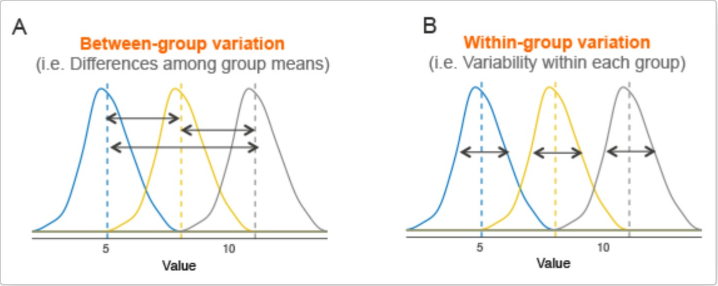
  versus

$$H_A : \text{Not all } \mu_j \text{ are the same} \quad \text{"There is a treatment effect"}$$

- Assumptions:
  1. Normal populations.
  2. Equality of population variances.

One-Way ANOVA: Introduction

- Assume that we have 3 groups to compare, as illustrated in the image below.



- The dashed line indicates the group mean.

- The idea behind the ANOVA test is very simple: if the average variation between groups is large enough compared to the average variation within groups, then you could conclude that at least one group mean is not equal to the others

One-Way ANOVA: Introduction

- Notation:

Treatment (level)	Observations				Averages
1	y_{11}	y_{12}	$\cdots$	y_{1n_1}	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	$\cdots$	y_{2n_2}	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kn_k}	$\bar{y}_{k\cdot}$

- ANOVA Table:

Source of Variation	*SS*	*df*	*MS*	*E{MS}*
Between treatments	$SSTR = \sum n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$k-1$	$MSTR = \frac{SSTR}{k-1}$	$\sigma^2 + \frac{\sum n_i(\mu_i - \mu_\cdot)^2}{k-1}$
Error (within treatments)	$SSE = \sum\sum(Y_{ij} - \bar{Y}_{i\cdot})^2$	$n_T - k$	$MSE = \frac{SSE}{n_T - k}$	σ^2
Total	$SSTO = \sum\sum(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$	$n_T - 1$		

One-Way ANOVA: Test Statistic

- The one-way ANOVA uses an F test statistic.

$$F_{cal} = \frac{MSTR}{MSE} \overset{under\ H_0}{\sim} F_{(k-1, n_T-k)}.$$

- H_0 is rejected if $F_{cal} > F^{1-\alpha}_{(k-1, n_T-k)}$ OR p-value $= P(F_{(k-1, n_T-k)} > F_{cal}) < \alpha$.

- Shortcut Formulae:
  - $SSTO = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}^2 - \frac{1}{n_T} \left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} \right)^2.$
  - $SSTR = \sum_{i=1}^{k} n_i \bar{Y}_{i\cdot}^2 - \frac{1}{n_T} \left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} \right)^2.$
  - $SSE = SSTO - SSTR.$

One-Way ANOVA: Example

Example

An economist compiled data on productivity improvements last year for a sample of firms producing electronic computing equipment. The firms were classified according to the level of their average expenditures for research and development in the past three years (low, moderate, high). The results of the study follow (productivity improvement is measured on a scale from 0 to 100). Assume that ANOVA model with the usual assumptions is appropriate.

		j											
i		1	2	3	4	5	6	7	8	9	10	11	12
1	Low	7.6	8.2	6.8	5.8	6.9	6.6	6.3	7.7	6.0			
2	Moderate	6.7	8.1	9.4	8.6	7.8	7.7	8.9	7.9	8.3	8.7	7.1	8.4
3	High	8.5	9.7	10.1	7.8	9.6	9.5						

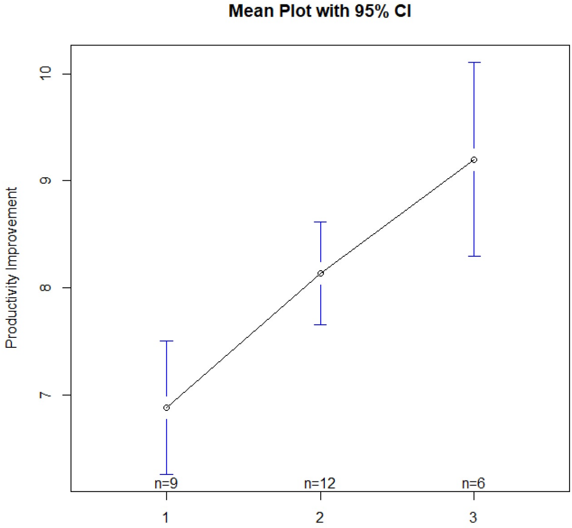
One-Way ANOVA: In R

```
> low <- c(7.6,8.2,6.8,5.8,6.9,6.6,6.3,7.7,6.0)
> moderate <- c(6.7,8.1,9.4,8.6,7.8,7.7,8.9,7.9,8.3,8.7,7.1,8.4)
> high <- c(8.5,9.7,10.1,7.8,9.6,9.5)
>
> prod.imp <- c(low,moderate,high)
>
> budget <- c(rep(1,9),rep(2,12),rep(3,6))
> budget <- factor(budget)
> results <- aov(prod.imp~budget)
> anova(results)
Analysis of Variance Table

Response: prod.imp
          Df Sum Sq Mean Sq F value    Pr(>F)
budget     2 20.125 10.0626   15.72 4.331e-05 ***
Residuals 24 15.362  0.6401
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

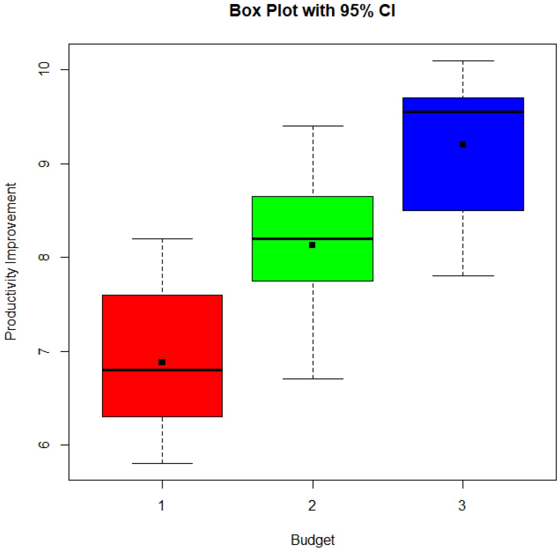
One-Way ANOVA: In R

```
> library(gplots)
> plotmeans(prod.imp ~ budget, xlab="Budget", ylab="Productivity Improvement",
+ main="Mean Plot with 95% CI")
```



One-Way ANOVA: In R

```
> means <- round(tapply(prod.imp, budget, mean),2)
> boxplot(prod.imp ~ budget, xlab="Budget", ylab="Productivity Improvement",
+ main="Box Plot with 95% CI", col=rainbow(3))
> points(means, col="black", pch=15)
```



One-Way ANOVA: Multiple Comparisons

- To determine which groups are different from the others **we need to conduct a POST HOC TEST** or a post hoc pair comparison.
- There are many post hoc tests available for analysis of variance.
- Let us use the **Tukey post hoc test**.
- The Tukey multiple comparison confidence limits for all pairwise comparisons $D = \mu_i - \mu_{i'}$ with family confidence coefficient of at least $1 - \alpha$ are as follows:

$$\hat{D} \pm T s_{\hat{D}},$$

  where

  - $\hat{D} = \bar{Y}_{i.} - \bar{Y}_{i'.}$.
  - $s^2_{\hat{D}} = MSE\left(\frac{1}{n_i} + \frac{1}{i'}\right)$.
  - $T = \frac{1}{\sqrt{2}} q_{(1-\alpha;k,n_T-k)}$. q is a critical value from the studentized range distribution.

One-Way ANOVA: Multiple Comparisons

- We wish to conduct a family of tests of the form

$$H_0 : \mu_i - \mu_{i'} = 0 \quad \text{versus} \quad H_A : \mu_i - \mu_{i'} \neq 0.$$

- In R:

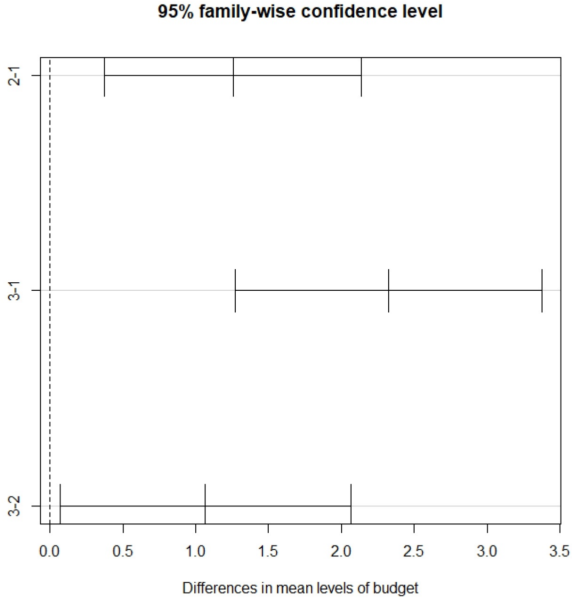
```
> tuk <- TukeyHSD(results, conf.level = 0.95)
> tuk
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = prod.imp ~ budget)

$budget
        diff        lwr      upr     p adj
2-1 1.255556 0.37453174 2.136579 0.0043755
3-1 2.322222 1.26919735 3.375247 0.0000335
3-2 1.066667 0.06767956 2.065654 0.0347870

> plot(tuk)
```

One-Way ANOVA: Multiple Comparisons



One-Way ANOVA: Diagnostic Checking

- Shapiro-Wilk test for normality:

  H_0 : The sample observations are taken from a Normal distribution.

- In R:

```
> res <- resid(results)
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.97377, p-value = 0.7033
```

- As the *p*-value is higher than the level of significance, you cannot reject the null hypothesis, which implies that the samples are taken from the normal populations.

One-Way ANOVA: Diagnostic Checking

- Another assumption requirement is the homogeneity of variances across the groups.

  H_0 : Equal variances across the cross-sectional group.

- Bartlett test for homogeneity is considered.

- In R:

```
> bartlett.test(prod.imp ~ budget)

        Bartlett test of homogeneity of variances

data:  prod.imp by budget
Bartlett's K-squared = 0.12936, df = 2, p-value = 0.9374

> vars <- round(tapply(prod.imp, budget, var),4)
> vars
     1      2      3
0.6619 0.5733 0.7520
```

- As the *p*-value is higher than the level of significance, we cannot reject the null hypothesis of homogeneity of variances across the three groups.

Two-Way ANOVA: Introduction

- Two-way ANOVA test is used to evaluate simultaneously the effect of two grouping variables (A and B) on a response variable.

Factor B / Factor A	(Level 1)	(Level 2)	...	(Level b)	(Mean)
(Level 1)	$Y_{111}, \ldots, Y_{11k}$	$Y_{121}, \ldots, Y_{12k}$	...	$Y_{1b1}, \ldots, Y_{1bk}$	$\overline{Y}_{1..}$
(Level 2)	$Y_{211}, \ldots, Y_{21k}$	$Y_{221}, \ldots, Y_{22k}$	...	$Y_{2b1}, \ldots, Y_{2bk}$	$\overline{Y}_{2..}$
	⋮	⋮	⋮	⋮	⋮
(Level a)	$Y_{a11}, \ldots, Y_{a1k}$	$Y_{a21}, \ldots, Y_{a2k}$	...	$Y_{ab1}, \ldots, Y_{abk}$	$\overline{Y}_{a..}$
(Mean)	$\overline{Y}_{.1.}$	$\overline{Y}_{.2.}$	...	$\overline{Y}_{.b.}$	$\overline{Y}_{...}$

Two-Way ANOVA: Introduction

- Two-way ANOVA test hypotheses:
  1. There is no difference in the means of factor *A*.
  $$H_0 : \mu_{1\cdot} = \mu_{2\cdot} = \ldots = \mu_{a\cdot}.$$
  2. There is no difference in means of factor *B*.
  $$H_0 : \mu_{\cdot 1} = \mu_{\cdot 2} = \ldots = \mu_{\cdot b}.$$
  3. There is no interaction between factors *A* and *B*.
- Assumptions of two-way ANOVA test:
  1. Observations within each cell are **normally distributed**.
  2. Observations within each cell have **equal variances**.

Two-Way ANOVA: ANOVA Table

Source of Variation	SS	df	MS	E{MS}
Factor A	$SSA = nb\sum(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2$	$a-1$	$MSA = \frac{SSA}{a-1}$	$\sigma^2 + bn\frac{\sum(\mu_{i\cdot} - \mu_{\cdot\cdot})^2}{a-1}$
Factor B	$SSB = na\sum(\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2$	$b-1$	$MSB = \frac{SSB}{b-1}$	$\sigma^2 + an\frac{\sum(\mu_{\cdot j} - \mu_{\cdot\cdot})^2}{b-1}$
AB interactions	$SSAB = n\sum\sum(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdot\cdot\cdot})^2$	$(a-1)(b-1)$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$\sigma^2 + n\frac{\sum\sum(\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot})^2}{(a-1)(b-1)}$
Error	$SSE = \sum\sum\sum(Y_{ijk} - \bar{Y}_{ij\cdot})^2$	$ab(n-1)$	$MSE = \frac{SSE}{ab(n-1)}$	σ^2
Total	$SSTO = \sum\sum\sum(Y_{ijk} - \bar{Y}_{\cdot\cdot\cdot})^2$	$nab-1$		

- Test Statistics
  1. Test for factor A main effects
  $$F_{cal} = \frac{MSA}{MSE} \sim F_{(a-1, ab(n-1))}.$$
  2. Test for factor A main effects
  $$F_{cal} = \frac{MSB}{MSE} \sim F_{(b-1, ab(n-1))}.$$
  3. Test for interaction
  $$F_{cal} = \frac{MSAB}{MSE} \sim F_{((a-1)(b-1), ab(n-1))}.$$

Two-Way ANOVA: Example

Example

The effective life (in hours) of batteries is compared by material type (1, 2 or 3) and operating temperature: Low (-10°C), Medium (20°C) or High (45°C). Twelve batteries are randomly selected from each material type and are then randomly allocated to each temperature level. The resulting life of all 36 batteries is shown below:

Life (in hours) of batteries by material type and temperature

		Temperature (˚C)		
		Low (-10˚C)	Medium (20˚C)	High (45˚C)
Material type	1	130, 155, 74, 180	34, 40, 80, 75	20, 70, 82, 58
	2	150, 188, 159, 126	136, 122, 106, 115	25, 70, 58, 45
	3	138, 110, 168, 160	174, 120, 150, 139	96, 104, 82, 60

- This example has two factors (material type and temperature), each with 3 levels.
- **Research question**: Is there difference in mean life of the batteries for differing material type and operating temperature levels?

Two-Way ANOVA: In R

```
> Y <- c(130,155,74,180,34,40,80,75,20,70,82,58,150,188,159,126,136,
+ 122,106,115,25,70,58,45,138,110,168,160,174,120,150,139,96,104,82,60)
>
> A <- c(rep("Type1", 12), rep("Type2", 12), rep("Type3", 12))
> A
 [1] "Type1" "Type1" "Type1" "Type1" "Type1" "Type1" "Type1" "Type1" "Type1" "Type1" "Type1" "Type1"
[13] "Type2" "Type2" "Type2" "Type2" "Type2" "Type2" "Type2" "Type2" "Type2" "Type2" "Type2" "Type2"
[25] "Type3" "Type3" "Type3" "Type3" "Type3" "Type3" "Type3" "Type3" "Type3" "Type3" "Type3" "Type3"
>
> B <- rep(c(rep("Low", 4), rep("Medium", 4), rep("High", 4)),3)
> B
 [1] "Low"    "Low"    "Low"    "Low"    "Medium" "Medium" "Medium" "Medium" "High"   "High"   "High"
[12] "High"   "Low"    "Low"    "Low"    "Low"    "Medium" "Medium" "Medium" "Medium" "High"   "High"
[23] "High"   "High"   "Low"    "Low"    "Low"    "Low"    "Medium" "Medium" "Medium" "Medium" "High"
[34] "High"   "High"   "High"
>
> data.frame(A, B, Y)
       A      B   Y
1  Type1    Low 130
2  Type1    Low 155
3  Type1    Low  74
4  Type1    Low 180
5  Type1 Medium  34
6  Type1 Medium  40
7  Type1 Medium  80
8  Type1 Medium  75
9  Type1   High  20
10 Type1   High  70
11 Type1   High  82
12 Type1   High  58
```

Two-Way ANOVA: In R

```
> fit1 <- aov(Y ~ A + B)
> anova(fit1)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
A          2  10684  5341.9  5.9472 0.006515 **
B          2  39119 19559.4 21.7759 1.239e-06 ***
Residuals 31  27845   898.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> means1 <- model.tables(fit1, type = 'means')
>
> means1
Tables of means
Grand mean
105.5278

A
 Type1  Type2  Type3
 83.17 108.33 125.08

B
  High    Low Medium
 64.17 144.83 107.58
```

Two-Way ANOVA: In R

```
> fit2 <- aov(Y ~ A * B)
> anova(fit2)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
A          2  10684  5341.9  7.9114  0.001976 **
B          2  39119 19559.4 28.9677 1.909e-07 ***
A:B        4   9614  2403.4  3.5595  0.018611 *
Residuals 27  18231   675.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> means2 <- model.tables(fit2, type = 'means')
> means2
  Tables of means
  Grand mean

  105.5278

  A
   Type1  Type2  Type3
   83.17 108.33 125.08

  B
    High    Low Medium
   64.17 144.83 107.58

 A:B
       B
A       High   Low    Medium
  Type1  57.50 134.75  57.25
  Type2  49.50 155.75 119.75
  Type3  85.50 144.00 145.75
```

Two-Way ANOVA: In R

- The ANOVA table gives:

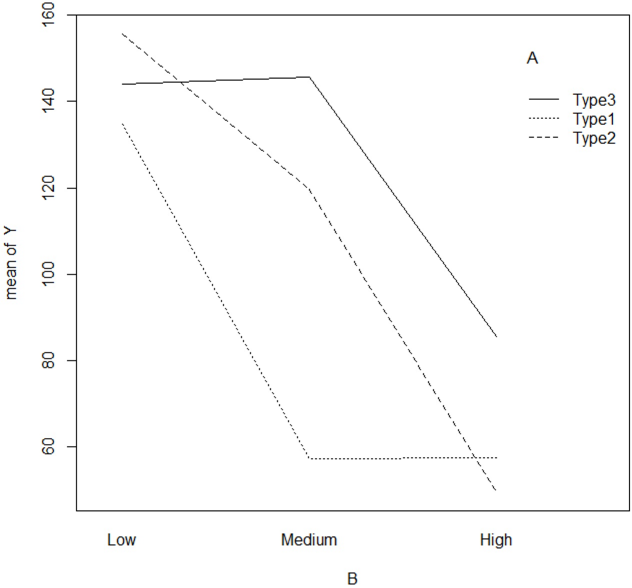
$$(F_{cal}, pvalue) = \{(7.91, 0.002), (28.97, < 0.0001), (3.56, 0.019)\},$$

  for material, operating temperature and material*temperature, respectively. So, both material and temperature are needed, as well as their interaction, to explain battery life.

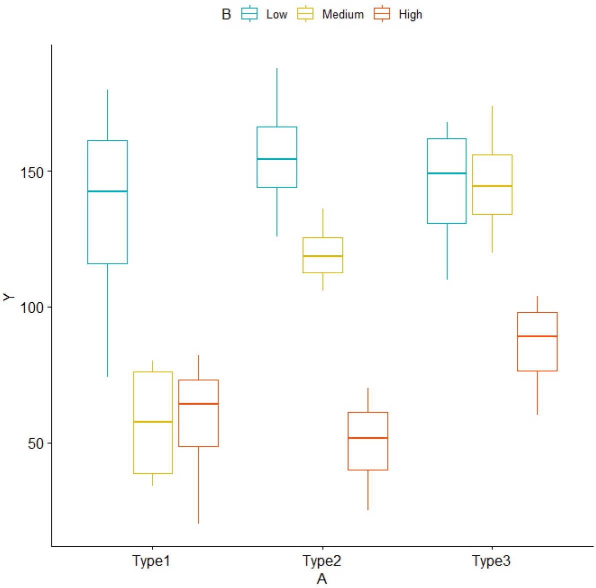
- It can be seen that, overall, battery life decreases with higher operating temperature, although battery life remains high for material 3 at medium temperature.

- Since the lines representing the three materials in the plot are not parallel, this implies there is an interaction effect between material and operating temperature.

Two-Way ANOVA: In R



Two-Way ANOVA: In R



Special Symbols Used in R Formula

Special symbols used in R formulas

Symbol	Usage
~	Separates response variables on the left from the explanatory variables on the right. For example, a prediction of `y` from A, B, and C would be coded `y ~ A + B + C`.
+	Separates explanatory variables.
:	Denotes an interaction between variables. A prediction of `y` from A, B, and the interaction between A and B would be coded `y ~ A + B + A:B`.
*	Denotes the complete crossing variables. The code `y ~ A*B*C` expands to `y ~ A + B + C + A:B + A:C + B:C + A:B:C`.
^	Denotes crossing to a specified degree. The code `y ~ (A+B+C)^2` expands to `y ~ A + B + C + A:B + A:C + A:B`.
.	A place holder for all other variables in the data frame except the dependent variable. For example, if a data frame contained the variables `y`, `A`, `B`, and `C`, then the code `y ~ .` would expand to `y ~ A + B + C`.

Other Designs

Formulas for common research designs

Design	Formula
One-way ANOVA	`y ~ A`
One-way ANCOVA with one covariate	`y ~ x + A`
Two-way Factorial ANOVA	`y ~ A * B`
Two-way Factorial ANCOVA with two covariates	`y ~ x1 + x2 + A * B`
Randomized Block	`y ~ B + A` (where B is a blocking factor)
Repeated measures ANOVA with one within-groups factor (W) and one between-groups factor (B)	`y ~ B * W + Error(Subject/W)`