

# Modul IS240 Lab 8: ANOVA

Prodi Sistem Informasi

## 1 Pendahuluan

### 1.1 Tujuan Pembelajaran

- Mahasiswa mampu melakukan estimasi dan inferensi rata-rata populasi dari sampel.

### 1.2 Topik Praktikum

- ANOVA

### 1.3 Petunjuk Umum

Gunakan file tutorial/ cheat sheet sebagai bahan referensi anda. File-file tersebut dapat diunduh dari Dropbox folder TutorialManual atau dari elearning.

- 2RTutorialTTH2018DescriptiveStatistics.pdf,

- 2RTutorialTTHHypothesisTesting.pdf

## 2 Materi Praktikum (Baca sebelum mengerjakan soal.)

Uji ANOVA digunakan bila kita ingin membandingkan mean beberapa populasi (3 atau lebih populasi). Tipe uji ANOVA yang dibahas dalam praktikum ini adalah One-Way ANOVA. Tipe ANOVA RBD dan two-way ANOVA di luar lingkup mata kuliah ini.

### 2.1 Syarat penggunaan uji ANOVA:

- data masing-masing populasi mengikuti distribusi Gaussian.
- setiap sampel diambil secara acak (random) dan independen.
- varians masing-masing populasi “*sama*”.

### 2.2 Setup uji hipotesa untuk ANOVA:

$$H_0 : \mu_A = \mu_B = \dots = \mu_k$$

$H_a$  : Minimal salah satu mean populasi tidak sama

Kriteria penolakan hipotesa null:

1. p-value < tingkat signifikansi alpha atau
2. test statistic F > nilai kritis dari tabel F.

Uji ANOVA meringankan pekerjaan kita. Bila pada materi sebelumnya kita membandingkan mean 2 populasi, maka dengan ANOVA kita tidak perlu membandingkan mean 2 populasi berkali-kali ( $\mu_A$  vs  $\mu_B$ ,  $\mu_A$  vs  $\mu_C$  dan seterusnya).

Bila hipotesa kesamaan mean ditolak, barulah kita akan mencari mean populasi mana yang berbeda.

## 2.3 Cara untuk mencari mean mana yang berbeda:

- informal: membandingkan mean secara visual dengan boxplot
- uji Tukey HSD (Honestly Significant Difference) bila ukuran sampel sama
- uji Tukey-Kramer bila ukuran sampel tidak sama.

## 2.4 Langkah-langkah uji ANOVA:

1. Cek asumsi ANOVA untuk data.
  - a. Uji normalitas dengan uji Shapiro-Wilk atau Anderson-Darling.
  - b. Cek kesamaan varians dengan boxplot atau menghitung varians atau dengan uji hipotesa kesamaan varians Levene.
2. Lakukan uji hipotesa kesamaan mean.
  - a. Bila  $H_0$  tidak dapat ditolak, stop dan simpulkan bahwa tidak ada perbedaan yang signifikan pada semua mean populasi.
  - b. Bila  $H_0$  ditolak, lakukan uji Tukey HSD atau Tukey-Kramer.

## 2.5 Contoh Black halaman 413

Misalnya kita ingin mengetahui apakah terdapat perbedaan pada rata-rata (mean) usia pekerja di 3 lokasi pabrik.

### 2.5.1 Impor data

```
age <- read.table(header=TRUE,text=" plant1 plant2 plant3
29 32 25
27 33 24
30 31 24
27 34 25
28 30 26
")
```

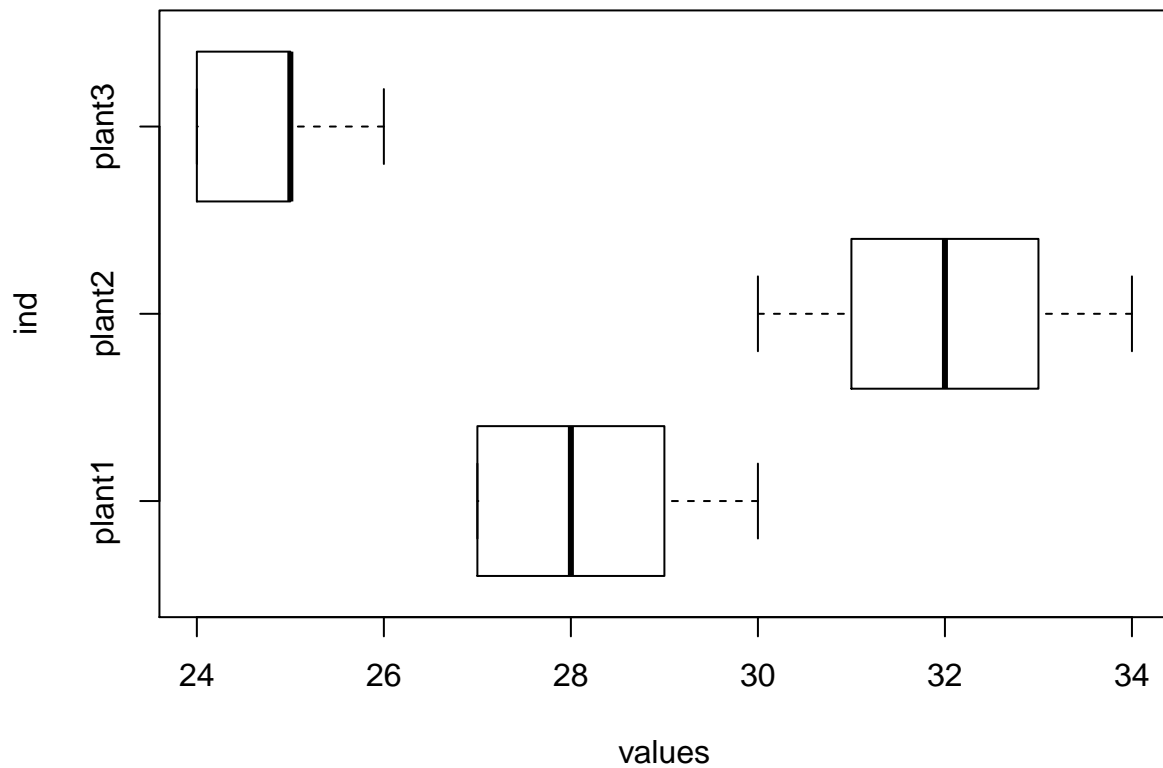
### 2.5.2 Statistika Deskriptif

```
#hitung mean usia pekerja untuk masing-masing plant
apply(age, 2, mean)
```

```
## plant1 plant2 plant3
##  28.2   32.0   24.8
```

```
#stack data agar R bisa memisahkan data berdasarkan lokasi plant.
s.age <- stack(age)
```

```
boxplot(values ~ ind, data = s.age, horizontal = TRUE)
```



### 2.5.3 Cek asumsi

Sebelum melakukan uji ANOVA, lakukan uji normalitas dan uji kesamaan varians.

*#uji normalitas*

```
apply(age, 2, shapiro.test)
```

```
## $plant1
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.90202, p-value = 0.4211
##
##
## $plant2
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.98676, p-value = 0.9672
##
##
## $plant3
```

```
##
## Shapiro-Wilk normality test
##
## data: newX[, i]
## W = 0.88104, p-value = 0.314
```

```
#hitung varians
apply(age, 2, var)
```

```
## plant1 plant2 plant3
## 1.7 2.5 0.7
```

```
#hitung cv
cv <- function(x){
  cv <- sd(x)/mean(x)
  return (cv)
}
apply(age, 2, cv)
```

```
## plant1 plant2 plant3
## 0.04623548 0.04941059 0.03373629
```

```
library(car)
leveneTest(s.age$values~s.age$ind)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 2 0.9333 0.42
##      12
```

- Hasil uji normalitas menunjukkan semua sampel mengikuti distribusi Gaussian.
- Varians usia pada lokasi 1, 2, dan 3 sepertinya tidak sama tetapi nilai CV hampir sama.
- Uji kesamaan varians Levene memberi kesimpulan tidak terdapat perbedaan signifikan antara varians ketiga populasi.

## 2.5.4 ANOVA

Gunakan fungsi `aov()` atau `lm()` untuk melakukan uji ANOVA.

```
#stack data agar R bisa memisahkan data berdasarkan lokasi plant.
s.age <- stack(age)
```

```
#uji ANOVA dengan fungsi aov
results <- aov(s.age$values ~ s.age$ind)
summary(results)
```

```
##      Df Sum Sq Mean Sq F value    Pr(>F)
## s.age$ind    2  129.7    64.87   39.71 5.11e-06 ***
## Residuals   12   19.6     1.63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tabel ANOVA menunjukkan nilai p-value yang sangat kecil  $< 5\%$ . Kecil peluangnya kita mendapatkan mean sampel usia pekerja pada ketiga lokasi ini seperti ini atau lebih buruk bila hipotesa null benar pada tingkat signifikansi 5%. Pada contoh ini hipotesa null kita tolak pada tingkat signifikansi 5%. Kita simpulkan minimal terdapat satu mean populasi yang berbeda.

Kita juga bisa membandingkan test statistic dengan nilai kritis. Pada buku Black nilai kritis distribusi F tercantum dalam tabel A7. Analisa varians menggunakan nilai kritis upper tail.

```
qf(p= 0.05, df1 = 2, df2 = 12, lower.tail = FALSE)
```

```
## [1] 3.885294
```

```
#atau
```

```
qf(p= 0.95, df1 = 2, df2 = 12, lower.tail = TRUE)
```

```
## [1] 3.885294
```

Test statistic > nilai kritis pada tingkat signifikansi 5%. Kita simpulkan bahwa minimal terdapat 1 mean yang populasi yang berbeda.

### 2.5.5 Tukey HSD

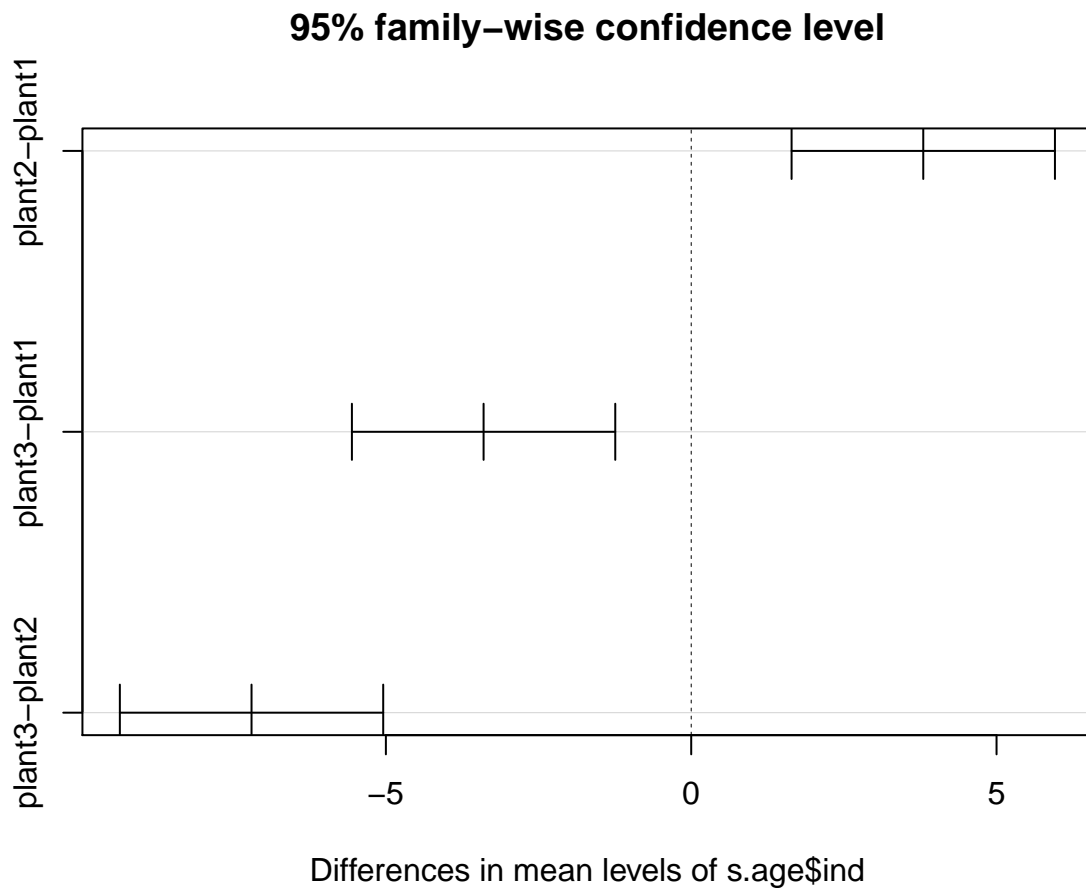
Setelah menolak hipotesa null, kita melakukan uji Tukey HSD untuk menentukan mean mana yang berbeda.

```
tk <- TukeyHSD(results, conf.level = .95)
```

```
tk
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = s.age$values ~ s.age$ind)
##
## $`s.age$ind`
##          diff          lwr          upr      p adj
## plant2-plant1  3.8  1.643591  5.956409 0.0013758
## plant3-plant1 -3.4 -5.556409 -1.243591 0.0032247
## plant3-plant2 -7.2 -9.356409 -5.043591 0.0000034
```

```
plot(tk)
```



Selang kepercayaan untuk semua selisih tidak mengandung 0 dan p-valuenya sangat kecil. Hasil uji Tukey HSD menunjukkan terdapat perbedaan yang signifikan antara mean usia pekerja masing-masing plant.

## 2.6 Contoh Black halaman 415 #11.6

### 2.6.1 Impor data

```
library(readxl)
dat <- read_excel("ANOVA_data.xlsx", sheet = "11_6")
dat
```

```
## # A tibble: 6 x 5
##   `1`   `2`   `3`   `4`   `5`
##   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1    14    10    11    16    14
## 2    13     9    12    17    12
## 3    10    12    13    14    13
## 4    NA     9    12    16    13
## 5    NA    10   NA    17    12
## 6    NA    NA    NA    NA    14
```

### 2.6.2 Cek asumsi

```
apply(dat, 2, mean, na.rm = TRUE) #hitung mean
```

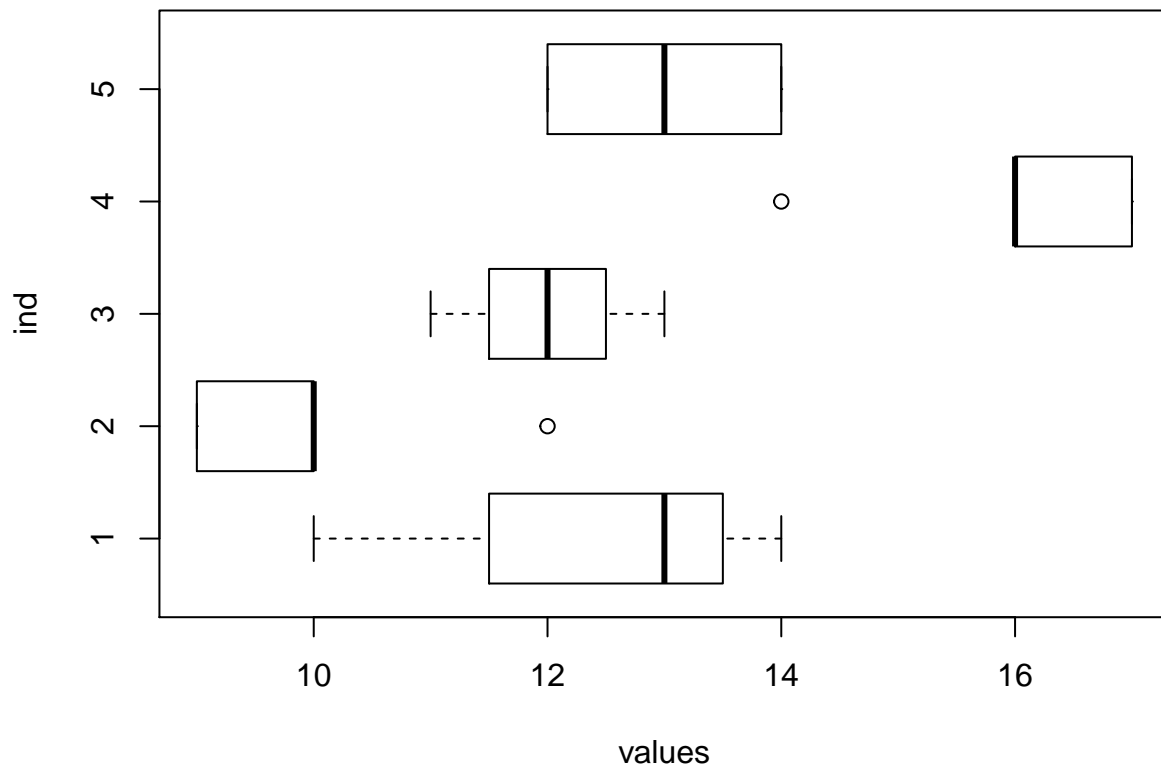
```
##      1      2      3      4      5
## 12.33333 10.00000 12.00000 16.00000 13.00000
```

```
apply(dat, 2, var, na.rm = TRUE) #hitung varians
```

```
##      1      2      3      4      5
## 4.333333 1.500000 0.6666667 1.500000 0.8000000
```

```
s.dat <- stack(dat)
```

```
boxplot(values ~ ind, s.dat, horizontal = TRUE)
```



```
apply(dat, 2, shapiro.test) #cek normalitas
```

```
## $`1`
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.92308, p-value = 0.4633
##
##
## $`2`
```

```
##
## Shapiro-Wilk normality test
##
## data: newX[, i]
## W = 0.83274, p-value = 0.1458
##
##
## $`3`
##
## Shapiro-Wilk normality test
##
## data: newX[, i]
## W = 0.94466, p-value = 0.683
##
##
## $`4`
##
## Shapiro-Wilk normality test
##
## data: newX[, i]
## W = 0.83274, p-value = 0.1458
##
##
## $`5`
##
## Shapiro-Wilk normality test
##
## data: newX[, i]
## W = 0.85319, p-value = 0.167
```

```
library(car)
car::leveneTest(values ~ ind, data = s.dat) #cek varians
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 4  0.4689 0.7578
##      18
```

- Dari nilai mean dapat diduga terdapat minimal 1 mean yang berbeda.
- Uji normalitas menunjukkan data kelima sampel Gaussian.
- Uji Levene menunjukkan kesamaan varians.

### 2.6.3 Uji ANOVA

```
results <- aov(values ~ ind, data = s.dat)
summary(results)
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## ind           4  93.77   23.442    15.82 1.02e-05 ***
## Residuals    18  26.67    1.481
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 7 observations deleted due to missingness
```

```
qf(0.01, df1 = 4, df2 = 18, lower.tail = FALSE)
```



```
## [1] 4.579036
```

- Test statistik > nilai kritis F dengan df 4 dan 18 dan alfa 1%.
- Hipotesa null ditolak pada tingkat signifikansi 1%.
- Kesimpulan: Minimal 1 mean populasi berbeda.

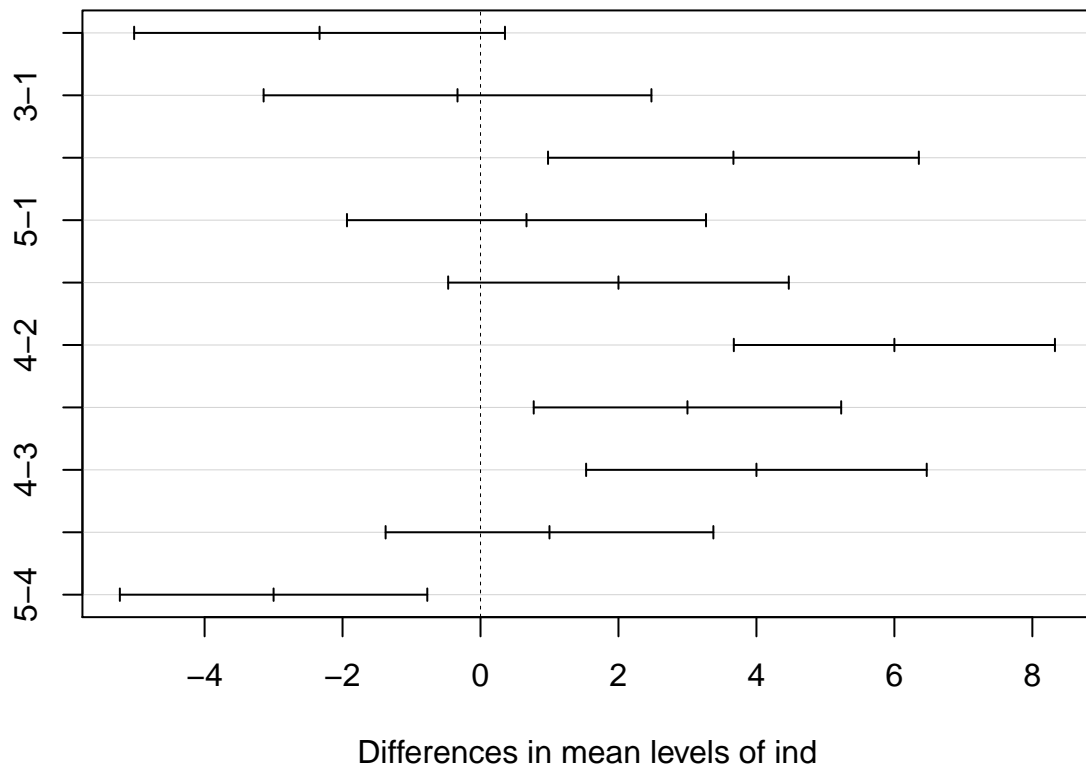
#### 2.6.4 Tukey-Kramer

```
tk <- TukeyHSD(results)
tk
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = values ~ ind, data = s.dat)
##
## $ind
##          diff          lwr          upr      p adj
## 2-1 -2.3333333 -5.0211518  0.3544851 0.1073308
## 3-1 -0.3333333 -3.1443212  2.4776545 0.9961302
## 4-1  3.6666667  0.9788482  6.3544851 0.0050403
## 5-1  0.6666667 -1.9358024  3.2691357 0.9346880
## 3-2  2.0000000 -0.4689189  4.4689189 0.1471533
## 4-2  6.0000000  3.6722809  8.3277191 0.0000032
## 5-2  3.0000000  0.7713787  5.2286213 0.0056640
## 4-3  4.0000000  1.5310811  6.4689189 0.0009654
## 5-3  1.0000000 -1.3757183  3.3757183 0.7103194
## 5-4 -3.0000000 -5.2286213 -0.7713787 0.0056640
```

```
plot(tk)
```

### 95% family-wise confidence level



## 3 SARAN

Bila anda tidak bisa menemukan  $F_{\alpha, df1, df2}$  pada tabel F, gunakan nilai terdekat dengan df terdekat atau gunakan  $1/F_{1-\alpha, df2, df1}$ .

```
qf(0.01, df1 = 4, df2 = 18, lower.tail = FALSE)
```

```
## [1] 4.579036
```

```
1/qf(0.99, df1 = 18, df2 = 4, lower.tail = FALSE)
```

```
## [1] 4.579036
```