BAB X
ANALISIS VARIANS (ANOVA)

Capaian Pembelajaran

- Mahasiswa mengerti dan memahami serta mampu melakukan pengujian secara serentak untuk beberapa populasi (lebih dari dua populasi)
- Mahasiswa dapat menguji asumsi ANOVA menggunakan perangkat lunak R-Studio.
- Mahasiswa mampu melakukan diagnosa hasil ANOVA menggunakan perangkat lunak R-Studio.
- Mahasiswa mampu menginterpretasi hasil ANOVA.

10.1 Pengertian dan Asumsi

Jika sampel $n > 2$, akan diasumsikan bahwa ada n sampel dari populasi. Salah satu prosedur yang sangat umum digunakan untuk menangani pengujian *mean* dari populasi adalah analisis varians, atau ANOVA. Prosedur analisis-varian bergantung pada distribusi yang disebut distribusi-*F* karena dipakai untuk menguji lebih dari 2 sampel. Suatu variabel dikatakan berdistribusi F jika distribusinya berbentuk kurva miring ke kanan (Weiss, 2016).

Analisis varians tentu bukan teknik baru bagi pembaca yang telah mengikuti materi tentang teori regresi. Pendekatan analisis varians digunakan untuk mempartisi jumlah total kuadrat menjadi sebagian untuk regresi dan sebagian untuk kesalahan (Walpole *et al.*, 2015). Asumsi yang harus dipenuhi dalam ANOVA adalah normalitas, homogenitas, linearitas dan independensi.

Berbeda dengan regresi dimana variabel independennya bersifat numerik (kuantitatif), ANOVA memiliki variabel independen yang bersifat kategorikal (kualitatif). Dalam ANOVA, variabel indepen disebut sebagai faktor yang memiliki sejumlah level atau perlakuan (*treatment*). Diasumsikan bahwa terdapat *k* populasi yang saling bebas dan berdistribusi normal dengan mean $\mu_1, \mu_2, \ldots, \mu_k$ dan varians σ^2. Hipotesis yang diuji dalam ANOVA adalah

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$

H_1 : Minimal terdapat satu $\mu_k \neq 0$

Model ANOVA dapat dituliskan sebagai berikut:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad i = 1, 2, \ldots, I \text{ dan } j = 1, 2, \ldots, J$$

dimana y_{ij} menunjukkan observasi ke-*j* dari perlakuan ke-*i* dengan struktur data seperti pada Tabel 10.1. μ merupakan rata-rata total dari semua nilai rata-rata sampel ke-i (μ_i), yaitu $\mu_i = \mu + \alpha_i$ dimana

$\sum_{i=1}^{k}\alpha_i = 0$ dan $\mu = \frac{1}{k}\sum_{i=1}^{k}\mu_i$, α_i adalah efek dari perlakuan ke-*i*, ε_{ij} mengukur penyimpangan antara observasi ke-*j* dari sampel ke-*i* dari rata-rata perlakuan yang sesuai.

Tabel 10.1 Struktur data ANOVA

Perlakuan	**1**	**2**	**…**	***i***	**…**	***k***	**Total**
	y_{11}	y_{21}	⋯	y_{i1}	⋯	y_{k1}	
	y_{12}	y_{22}	⋯	y_{i2}	⋯	y_{k2}	
	⋮	⋮		⋮		⋮	
	y_{1n}	y_{2n}	⋯	y_{in}	⋯	y_{kn}	
Total	$Y_{1.}$	$Y_{2.}$	⋯	$Y_{i.}$	⋯	$Y_{k.}$	$Y..$
Rata-rata total	$\bar{y}_{1.}$	$\bar{y}_{2.}$	…	$\bar{y}_{i.}$	⋯	$\bar{y}_{k.}$	$\bar{y}..$

Menurut (Weiss, 2016), untuk mengatur dan meringkas jumlah yang diperlukan untuk melakukan analisis menggunakan *one way* ANOVA, dapat menggunakan tabel *one way* ANOVA. Format umum dari tabel tersebut dapat dilihat pada Tabel berikut

Tabel 10.1. ANOVA

Source	df	SS	MS = SS/df	*F-statistics*
Treatment	$k-1$	SSTR	MSTR = SSTR/k −1	F = MSTR/MSE
Error	$n-k$	SSE	MSE = SSE/n − k	
Total	$n-1$	SST		

dimana

Sum of Squares	*Defining formula*	*Computing formula*
Total, SST	$\sum_{i=1}^{n}(x_i-\bar{x})^2$	$\sum_{i=1}^{n}{x_i}^2 - {(\sum_{i=1}^{n}x_i)^2}/{n}$
Treatment, SSTR	$\sum_{i=1}^{n}n_j(\bar{x}_j-\bar{x})^2$	$\sum_{i=1}^{n}\left({T_j^2}/{n_j}\right) - {(\sum_{i=1}^{n}x_i)^2}/{n}$
Error, SSE	$\sum_{i=1}^{n}(n_j-1){s_j}^2$	*SST− SSTR*

10.2 Aplikasi Menggunakan R

Berikut adalah contoh data yang bersumber dari (Bruce, Bruce and Gedeck, 2020). Data tersebut menunjukkan pengunjung dari empat halaman web, yang didefinisikan sebagai jumlah pengunjung dihabiskan di halaman dalam detik. Keempat halaman tersebut dialihkan sehingga setiap pengunjung web menerima satu secara acak. Ada total lima pengunjung untuk setiap halaman, setiap kolom adalah kumpulan data yang independen. Pengunjung pertama untuk halaman 1 tidak memiliki koneksi ke

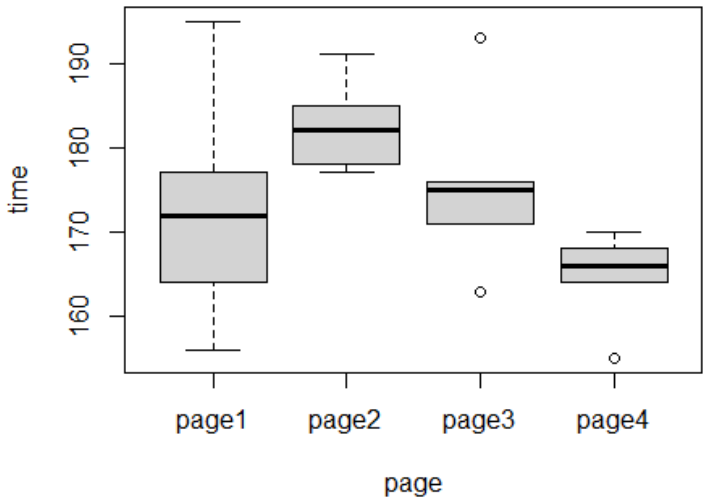
pengunjung pertama untuk halaman 2. Perhatikan bahwa dalam pengujian web seperti ini, tidak sepenuhnya menerapkan desain pengambilan sampel acak klasik di mana setiap pengunjung dipilih secara acak dari beberapa populasi besar. Kita harus memilih pengunjung pada saat mereka hadir. Pengunjung mungkin berbeda secara sistematis tergantung pada waktu, musim pada tahun ini, kondisi internet mereka, perangkat apa yang mereka gunakan, dan sebagainya. Faktor-faktor ini harus dianggap sebagai potensi bias ketika hasil percobaan ditinjau.

Tabel 10.2. Data

	Page 1	**Page 2**	**Page 3**	**Page 4**
	164	178	175	155
	172	191	193	166
	177	182	171	164
	156	185	163	170
	195	177	176	168
Rata-rata	172	185	176	162
Rata-rata total				173.75

Dengan empat rata-rata, ada enam kemungkinan perbandingan antar kelompok. Apakah semua halaman memiliki pengunjung yang sama, dan perbedaan di antara mereka disebabkan oleh keacakan di mana kumpulan waktu sesi yang sama dialokasikan di antara keempat halaman?. Perintah yang dapat anda tuliskan pada R adalah :

```
> library(lmPerm)
> dataku<-read.csv("dataku.csv")
> page<c(rep("page1",5),rep("page2",5),rep("page3",5),rep("page4",5))
> page
   [1] "page1" "page1" "page1" "page1" "page1" "page2" "page2"
   [8] "page2" "page2" "page2" "page3" "page3" "page3" "page3"
   [15] "page3" "page4" "page4" "page4" "page4" "page4"
> time<-c(dataku$page1,dataku$page2,dataku$page3,dataku$page4)
> time
   [1] 164 172 177 156 195 178 191 182 185 177 175 193 171 163 176
   [16] 155 166 164 170  168
> boxplot(time~page,data=df)
```




```
> summary(aovp(time ~ page, data=df))
  [1] "Settings:  unique SS "
  Component 1 :
              Df R Sum Sq R Mean Sq Iter Pr(Prob)
  page1        3    831.4    277.13 3104  0.09278 .
  Residuals   16   1618.4    101.15
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Nilai *p*, yang diberikan oleh `Pr(Prob)`, adalah 0,09278, yang berarti bahwa 9,3% dari tingkat waktu, respon pengunjung di antara empat halaman web mungkin berbeda sebanyak yang sebenarnya diamati, hanya secara kebetulan. Tingkat ketidakmungkinan ini jauh dari taraf 5%, kita dapat menggunakan taraf 10% untuk menyimpulkan perbedaan di antara empat halaman bisa saja muncul secara kebetulan.

10.3 Latihan

1. Data diambil dari majalah US *Motor Trend* 1974, dan terdiri dari konsumsi bahan bakar dan 10 aspek desain dan performa mobil untuk 32 mobil (model 1973-1974). *Mpg Miles*/(US) *gallon*. Cyl (*Number of cylinders*), disp (*Displacement* (cu.in.)), hp (*Gross horsepower*), drat (*Rear axle ratio*), wt (*Weight* (1000 lbs)), qsec (1/4 *mile time*), vs (*Engine* (0 = *V-shaped*, 1 = *straight*)), am (*Transmission* (0 = *automatic*, 1 = *manual*)), *gear* (*Number of forward gears*).

   Data dapat dipanggil dari R dengan mengetikkan `mtcars`. Lakukan pengujian asumsi, kemudian analisis data untuk melihat perbedaan rata-rata setiap variabel.

2. Diberikan data enam mesin berbeda yang dipertimbangkan untuk digunakan dalam manufaktur segel karet. Keenam mesin tersebut dibandingkan sehubungan dengan kekuatan tarik produk. Sampel acak dari empat segel dari setiap mesin digunakan untuk menentukan apakah kekuatan tarik rata-rata bervariasi dari masing-masing mesin. Berikut adalah data ukuran kekuatan tarik dalam kilogram per cm^2 x 10^{-1}.

1	**2**	**3**	**4**	**5**	**6**
17.5	16.4	20.3	14.6	17.5	18.3
16.9	19.2	15.7	16.7	19.2	16.2
15.8	17.7	17.8	20.8	16.5	17.5
18.6	15.4	18.9	18.9	20.5	20.1

Lakukan analisis varians pada tingkat signifikansi (α) 5% dan tentukan apakah terdapat perbedaan rata-rata kekuatan tarik dari keenam mesin tersebut.