Statistika Sains Data dengan R

Modul Praktikum dengan Software Opensource R

Edisi Pertama

Egi Safitri, S.Mat., M.Si
Bandar lampung, Indonesia

Program Studi Sains Data
Institut Informatika dan Bisnis Darmajaya
Bandar Lampung
2023

MODUL 7

Teknik Sampling

7.1 Definisi Sampling

Mari kita katakan bahwa kita memiliki suatu populasi dengan ukuran N, sebuah sampel tidak lebih dari suatu subset data yang diambil dari populasi tersebut. Proses pemilihan sampel dikenal sebagai pengambilan sampel.

Pengambilan sampel acak didefinisikan sebagai teknik statistik yang digunakan dalam proses pengambilan sampel, di mana setiap item dalam populasi memiliki peluang dan kemungkinan yang setara untuk dipilih dalam sampel.

Dalam konteks ini, seleksi item sepenuhnya tergantung pada keberuntungan atau probabilitas, dan itulah sebabnya teknik pengambilan sampel ini sering disebut sebagai metode peluang. Pengambilan sampel acak sederhana merupakan metode dasar dalam pengambilan sampel dan sering kali menjadi komponen integral dari metode pengambilan sampel yang lebih kompleks. Ciri utama dari metode ini adalah bahwa setiap sampel memiliki probabilitas yang sama untuk dipilih.

Ukuran sampel yang ideal dalam pengambilan sampel acak ini sebaiknya lebih dari beberapa ratus agar metode pengambilan sampel acak sederhana dapat diterapkan secara efektif. Meskipun dianggap secara teoritis sederhana, penerapannya dalam praktik seringkali dianggap sulit. Bekerja dengan ukuran sampel yang besar bukanlah tugas yang mudah, dan terkadang menemukan kerangka sampel yang realistis dapat menjadi tantangan tersendiri.

Peneliti memiliki beberapa metode untuk membuat sampel acak sederhana.

Salah satunya adalah metode lotere, di mana setiap anggota populasi diberi nomor dan nomor tersebut dipilih secara acak. Sebagai contoh, pengambilan sampel acak sederhana dapat dilakukan dengan memilih nama 25 karyawan dari sebuah perusahaan yang memiliki 250 karyawan. Dalam situasi ini, seluruh populasi terdiri dari 250 karyawan, dan sampel dianggap acak karena setiap karyawan memiliki peluang yang sama untuk dipilih. Pengambilan sampel acak ini sering digunakan dalam dunia ilmu pengetahuan untuk melakukan uji kontrol acak atau eksperimen buta.

Pengambilan sampel dapat digunakan dalam dua skenario di bawah ini:

1. Ketika seluruh data populasi tidak tersedia Dalam kasus ini, kita mungkin harus menggunakan sampel data yang tersedia untuk membuat inferensi tentang seluruh populasi.
2. Ketika data populasi terlalu besar
3. Dalam hal ini, kita mungkin menggunakan salah satu teknik pengambilan sampel di bawah ini untuk membuat sampel dari populasi, dan selanjutnya inferensi dapat dibuat.

Catatan: Penting untuk memastikan bahwa sampel yang ideal dipilih karena sampel yang salah dapat mengarah pada inferensi yang tidak berkorelasi dengan populasi.

Metode pengambilan sampel diklasifikasikan menjadi dua kategori utama:

1. Pengambilan Sampel Probabilistik
2. Pengambilan Sampel Non-Probabilistik
3. Pengambilan Sampel Probabilistik:

Dalam pengambilan sampel ini, sampel dipilih secara acak, dan setiap sampel memiliki probabilitas diketahui untuk dipilih.

Pengambilan Sampel Probabilistik lebih lanjut diklasifikasikan sebagai berikut:

1. Pengambilan Sampel Acak Sederhana
2. Pengambilan Sampel Sistematik
3. Pengambilan Sampel Berstrata
4. Pengambilan Sampel Klaster

7.1.1 Pengambilan Sampel Acak Sederhana

Pengambilan Sampel Acak merupakan salah satu metode pengambilan sampel yang paling populer dan sering digunakan. Dalam pengambilan sampel acak sederhana, setiap kasus dalam populasi memiliki probabilitas yang sama untuk terpilih dalam sampel.

Sampel acak dapat diperoleh dengan memberi label pada semua kasus secara berurutan dan menghasilkan angka acak yang seragam untuk memilih kasus dari populasi.

Sampel acak sederhana dalam R dapat dihasilkan seperti berikut menggunakan fungsi sample():

Fungsi sampel didefinisikan sebagai berikut:

```
sample(x, size, replace = FALSE, prob = NULL)
```

Contoh pengambilan sampel ukuran 10 dari angka 1 hingga 10 dapat dihasilkan seperti berikut.

```
sample(1:10,10)
```

```
[1]  8  2  9  10  3  7  6  5  1  4
```

Seperti yang terlihat pada contoh di atas, sampel secara default dihasilkan tanpa penggantian, artinya, suatu item yang sudah dipilih untuk disampel tidak akan digunakan lagi untuk disampel. Sampel dengan penggantian dapat dibuat dengan mengatur parameter replace menjadi TRUE seperti di bawah ini.

```
sample(1:10,replace=T)
```

```
[1]  8  8  8  9  3  5  3  2  10  7
```

Seperti yang terlihat pada sampel di atas, angka 8 muncul tiga kali dan angka 3 muncul dua kali dalam sampel.

Pada dasarnya, dalam pengambilan sampel acak, semua item memiliki probabilitas yang sama ($p = 0,1$ dalam kasus di atas) untuk terpilih. Namun, dalam metode `sample()`, kita juga dapat menetapkan probabilitas untuk item yang akan terpilih dalam suatu sampel.

Misalnya, kita memiliki daftar dengan 2 item (merah, hijau), secara default baik merah maupun hijau memiliki peluang 50% ($p = 0,5$) untuk terpilih. Misalkan kita membutuhkan lebih banyak item merah dalam sampel daripada

hijau, hal ini dapat dilakukan dengan menggunakan parameter 'prob' dalam metode sample() seperti di bawah ini.

```
1 sample(c("merah","hijau"),10,replace=T,prob=c(0.6,0.4)
    )
2 [1] "hijau" "hijau" "hijau" "merah" "hijau" "merah" "
    merah" "hijau" "hijau" "merah"
```

Dapat kita lihat bahwa tampaknya ada lebih banyak warna merah dalam sampel daripada hijau, karena probabilitas pemilihan warna merah ($p = 0,6$) diatur lebih tinggi daripada hijau ($p = 0,4$).

7.1.2 Sampling Sistematis

Pemilihan sistematis digunakan dalam situasi di mana data populasi tersusun dalam daftar berurutan atau diatur berdasarkan waktu. Sebagai contoh, untuk menganalisis rata-rata penjualan suatu toko pada hari Minggu, pemilihan sistematis dapat digunakan dengan memilih data rata-rata penjualan pada hari ke-7 (Minggu) dari minggu tersebut untuk dimasukkan dalam sampel.

Artinya, dalam pemilihan sistematis, individu dipilih pada interval tetap dari data populasi. Untuk membuat sampel ukuran n dari populasi ukuran p, interval tetap (k) diambil sebagai $\frac{p}{n}$.

Artinya, $k = \frac{p}{n}$.

Artinya, untuk populasi berukuran 1000, untuk membuat sampel ukuran 100 (1000/100), setiap item ke-10 dari titik awal acak dapat dipilih untuk dimasukkan dalam sampel.

Sekarang, mari kita lihat bagaimana membuat sampel sistematis di R,

Sampel sistematis di R:

Untuk membuat sampel sistematis di R, fungsi S.SY() dari paket "TeachingSampling" digunakan.

```
1 #install.packages("TeachingSampling")
2 library(TeachingSampling)
3 P <- c("Mon-8", "Tues-4", "Wed-4", "Thurs-6", "Fri-7",
    "Sat-45","Sun-34","Mon-21", "Tues-11","Wed-34","
    Thurs-16","Fri-10","Sat-17","Sun-19")
```

```
4 #systematic sample from a population of 14 with every
    2nd included from the populaion P
5 systematic_sample <- S.SY(14,2)
6 systematic_sample
7 P[systematic_sample]
8 > P[systematic_sample]
9 [1] "Mon-8"    "Wed-4"    "Fri-7"    "Sun-34"   "Tues
    -11"  "Thurs-16" "Sat-17"
```

Dengan menggunakan kode R di atas, dari populasi P yang berisi unit yang terjual setiap hari selama periode 14 hari, kita telah membuat sampel sistematis hanya dengan unit yang terjual setiap hari selang.

Catatan: Pemilihan sistematis lebih mudah dilaksanakan dibandingkan Pemilihan Acak Sederhana. Namun, dalam pemilihan sistematis, tidak setiap item memiliki peluang yang sama untuk dipilih, sehingga banyak item mungkin tidak pernah terpilih. Jika suatu populasi memiliki kecenderungan periodik, efektivitas sampel sistematis bergantung pada hubungan antara interval periodik dan interval pemilihan sistematis.

7.1.3 Pemilihan Sample Berstrata

Dalam pemilihan sampel berstrata, populasi dibagi menjadi kelompok-kelompok kecil berdasarkan faktor-faktor umum yang paling baik menggambarkan seluruh populasi, seperti usia, jenis kelamin, pendapatan, dll. Kelompok yang terbentuk disebut sebagai stratum/strata. Sebagai contoh, untuk menganalisis waktu yang dihabiskan oleh pengguna pria dan wanita dalam mengirim pesan per hari, strata bisa diambil sebagai pengguna pria dan wanita, dan pemilihan acak dapat digunakan untuk memilih item di dalam strata pria dan wanita.

Catatan: Pemilihan sampel berstrata memberikan perkiraan yang lebih tepat dibandingkan dengan pemilihan acak, tetapi kelemahan terbesarnya adalah bahwa ini memerlukan pengetahuan tentang karakteristik yang sesuai dari populasi (detailnya tidak selalu tersedia), dan bisa sulit untuk memutuskan karakteristik mana yang akan dijadikan dasar stratifikasi.

Pemilihan Sampel Berstrata di R

Menggunakan `dplyr`

Mari kita lihat bagaimana membuat sampel berstrata menggunakan dataset iris dengan 3 sampel dari setiap spesies.

```
library(dplyr)
set.seed(1)
iris %>%
group_by (Species) %>%
sample_n(., 3)

        Sepal.Length Sepal.Width Petal.Length Petal.
           Width    Species
1           4.3          3.0           1.1          0.1
       setosa
2           5.7          3.8           1.7          0.3
       setosa
3           5.2          3.5           1.5          0.2
       setosa
4           5.7          3.0           4.2          1.2
   versicolor
5           5.2          2.7           3.9          1.4
   versicolor
6           5.0          2.3           3.3          1.0
   versicolor
7           6.5          3.0           5.2          2.0
   virginica
8           6.4          2.8           5.6          2.2
   virginica
9           7.4          2.8           6.1          1.9
   virginica
```

Menggunakan `strata()`:

Sampel berstrata yang sama di atas juga dapat dibuat menggunakan fungsi strata dari paket pengambilan sampel seperti berikut:

```
library(sampling)
stratas = strata(iris, c("Species"),size = c(3,3,3),
   method = "srswor")
```

```
stratas
Species ID_unit   Prob Stratum
17       setosa    17 0.06       1
24       setosa    24 0.06       1
45       setosa    45 0.06       1
65   versicolor    65 0.06       2
79   versicolor    79 0.06       2
95   versicolor    95 0.06       2
114   virginica   114 0.06       3
116   virginica   116 0.06       3
128   virginica   128 0.06       3
```

Catatan: Dalam fungsi di atas, metode mewakili metode yang digunakan untuk memilih sampel individu dalam strata. Metode-metode berikut umumnya digunakan.

- **srswor**: pemilihan acak sederhana tanpa penggantian
- **srswr**: pemilihan acak sederhana dengan penggantian
- **poisson**: sampel Poisson
- **systematic**: pemilihan sistematis

7.1.4 Pemilihan Sample Berkelompok

Pemilihan sampel berkelompok umumnya digunakan dalam kasus di mana data populasi bersifat geografis atau ketika ada kelompok yang telah ditentukan dalam populasi berdasarkan demografi, kebiasaan, latar belakang, dll. Dalam pemilihan sampel berkelompok, populasi pertama-tama dibagi menjadi kelompok-kelompok kecil yang dikenal sebagai kelompok, dan kemudian kelompok-kelompok acak dipilih untuk membuat sampel.

Jika semua elemen dari kelompok yang dipilih dimasukkan dalam sampel, maka disebut sebagai pemilihan sampel berkelompok satu tahap (Single-stage cluster sampling), dan jika pemilihan acak elemen dari setiap kelompok dimasukkan dalam sampel, maka disebut sebagai pemilihan sampel berkelompok dua tahap (Two-stage cluster sampling).

Sebagai contoh, misalkan sebuah organisasi ingin menganalisis efek samping dari suatu obat di seluruh Amerika Serikat, dalam hal ini, pemilihan sampel berkelompok dua tahap dapat dilakukan dengan pertama-tama membagi seluruh populasi menjadi kota-kota (di mana data setiap kota berisi detail tentang efek samping obat untuk semua pasien) dan kemudian secara acak memilih pasien di dalam kota-kota ini untuk dimasukkan dalam sampel.

Pemilihan Sampel Berkelompok di R:

Untuk melakukan pemilihan sampel berkelompok, kami telah menggunakan dataset Beban Kerja Guru Sekolah Dasar dari paket SDaA. Dataset beban kerja berisi detail beban kerja seperti jam kerja, waktu persiapan, dll. dari guru-guru di berbagai sekolah di berbagai distrik.

```
1  install.packages("SDaA")
2  library(SDaA)
3  data("teachers")
4  > head(teachers)
5  dist school hrwork size preprmin assist
6  1 large      12   35.00    26       210       0
7  2 large      12   35.00    18        75       0
8  3 large      12   35.00    27       300       0
9  4 large      12   34.60    34        90       0
10 5 large      12   33.75    30       180       0
11 6 large      12   35.00    27       300       0
12 #list of all the school_ids
13 > unique(teachers[,2])
14 [1] 12 13 20 21 22 36 38 41 11 30 31 32  4 23  7 15 16
       28 29  6  2 18 19 33 34  1  8  9  3 24 25
15 #creating a cluster sample with 7 randomly selected
      clusters.
16 #Here we have formed clusters using the school
      variable.Hence each cluster contains the workload
      data of 7 randomly selected schools.
17 set.seed(123456)
18 cl=cluster(teachers,clustername=c("school"),size=7,
      method="srswor")
```

```
cl_data = getdata(teachers, cl)
> head(cl_data)
dist hrwork size preprmin assist school ID_unit
   Prob
260 sm/me  30.00    9       NA      0      1     260
   0.2258065
243 large  35.00   20      225    600     18     243
   0.2258065
244 large  35.00   16       90    300     18     244
   0.2258065
19  large  37.50   25      180      0     20      19
   0.2258065
18  large  38.75   24      240      0     20      18
   0.2258065
17  large  38.35   24      120      0     20      17
   0.2258065
#list of the randomly selected schools
> unique(cl_data[,6])
[1]  1 18 20 25 28 31 41
#count of workload details within each school clusters
> table(cl_data$school)
8 12 16 21 28 34 38
5 13 24  7 18  7 10
#random sampling of clusters with a sample size of 5,
   so that each cluster contains 5 randomly selected
   workload details per school cluster.
cl_sam <- cl_data %>% group_by(school) %>% sample_n(
   size = 5)
#Each of the 7 clusters have 5 randomly selected
   workload data.
> table(cl_sam$school)
8 12 16 21 28 34 38
5  5  5  5  5  5  5
```

Dengan menggunakan kode R di atas, kami telah membuat 7 kelompok acak

di mana setiap kelompok berisi data beban kerja sekolah tertentu. Kelompok-kelompok ini kemudian diambil sampel secara acak dengan ukuran sampel sebanyak 5. Oleh karena itu, setiap kelompok memiliki 5 data beban kerja untuk setiap sekolah yang terpilih.

7.1.5 Perbedaan antara Pemilihan Sampel Berstrata dan Sampel Klaster

Perbedaan utama antara pemilihan sampel berstrata dan sampel klaster adalah bahwa dalam sampel klaster, kelompok/klaster muncul secara alami seperti kota, distrik, dll, dan elemen-elemen klaster yang dipilih secara keseluruhan digunakan untuk pengambilan sampel, misalnya, untuk data beban kerja, awalnya dipilih 7 kelompok sekolah dan semua elemen dari kelompok-kelompok ini digunakan untuk pengambilan sampel lebih lanjut. Artinya, data beban kerja dari hanya 7 sekolah digunakan dalam pengambilan sampel, mengabaikan detail beban kerja dari sekolah lainnya.

Sementara itu, dalam pemilihan sampel berstrata, kelompok (strata dalam hal ini) tidak ada pada awalnya, dan elemen dari setiap strata yang dibuat dipilih untuk dimasukkan dalam sampel. Sebagai contoh, untuk iris di atas, 3 elemen dari setiap dari tiga spesies yang tersedia telah dimasukkan dalam sampel, dan tidak ada spesies yang diabaikan secara keseluruhan seperti dalam kasus Sampel Klaster.