

# **Modul Teknik Pengumpulan Data**

Muhammad Ammar Sahab

9/22/2022

## **Table of contents**

# Home

Modul Teknik Pengumpulan Data.

# 1 Review Statistika Dasar

Mengumpulkan data pada dasarnya berarti mencari nilai-nilai dari variabel tertentu yang menggambarkan suatu objek. **Skala pengukuran** apa saja yang bisa dipakai? Dalam kata lain, jenis variabel apa saja yang dapat menggambarkan objek?

## 1.0.1 Skala Nominal

Anggap Anda memiliki nilai data mengenai warna rambut dan mata sebagai berikut:

Hair	Eye	Sex	n
Black	Brown	Male	32
Brown	Brown	Male	53
Red	Brown	Male	10
Blond	Brown	Male	3
Black	Blue	Male	11
Brown	Blue	Male	50
Red	Blue	Male	10
Blond	Blue	Male	30
Black	Hazel	Male	10
Brown	Hazel	Male	25
Red	Hazel	Male	7
Blond	Hazel	Male	5
Black	Green	Male	3
Brown	Green	Male	15
Red	Green	Male	7
Blond	Green	Male	8
Black	Brown	Female	36
Brown	Brown	Female	66
Red	Brown	Female	16
Blond	Brown	Female	4
Black	Blue	Female	9
Brown	Blue	Female	34
Red	Blue	Female	7
Blond	Blue	Female	64
Black	Hazel	Female	5

	Hair	Eye	Sex	n
	Brown	Hazel	Female	29
	Red	Hazel	Female	7
	Blond	Hazel	Female	5
	Black	Green	Female	2
	Brown	Green	Female	14
	Red	Green	Female	7
	Blond	Green	Female	8

Skala tersebut disebut **skala nominal**. Skala nominal **mengelompokkan** observasi, tanpa mengurutkan. Skala tersebut berupa **kualitatif** sehingga tidak direpresentasikan angka. Dalam kasus ini, tidak ada urutan tertentu; tidak ada warna mata terbaik, atau warna rambut terbaik.

Contoh representasi skala nominal adalah *factor* di bahasa R:

```
1 library(tidyverse)

-- Attaching packages ----- tidyverse 1.3.2 --
v ggplot2 3.3.6      v purrr  0.3.4
v tibble  3.1.6      v dplyr  1.0.9
v tidyr   1.2.0      v stringr 1.4.0
v readr   2.1.2      v forcats 0.5.1
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()

1 HairEyeColor |>
2   tibble::as_tibble() |> summary()
```

Hair	Eye	Sex	n
Length:32	Length:32	Length:32	Min. : 2.00
Class :character	Class :character	Class :character	1st Qu.: 7.00
Mode :character	Mode :character	Mode :character	Median :10.00
			Mean :18.50
			3rd Qu.:29.25
			Max. :66.00

Representasi data tersebut masih sebuah karakter. Ubah representasi menjadi faktor:

```

1 library(tidyverse)
2
3 HairEyeColor |>
4   tibble::as_tibble() |>
5   mutate_if(is.character, as.factor) |>
6   summary()

```

Hair	Eye	Sex	n
Black:8	Blue :8	Female:16	Min. : 2.00
Blond:8	Brown:8	Male :16	1st Qu.: 7.00
Brown:8	Green:8		Median :10.00
Red :8	Hazel:8		Mean :18.50
			3rd Qu.:29.25
			Max. :66.00

Setelah mengubah karakter menjadi representasi faktor, kita mengetahui golongan-golongan rambut yang ada di dataset. Jika representasi skala tertentu di suatu *software* benar, *software* tersebut sering memiliki alat-alat untuk menangani skala dengan baik.

## 1.0.2 Skala Ordinal

Jika kita memasukkan *level* atau tingkatan dari faktor, dan mengurutkannya, kita membuat suatu variabel dengan skala ordinal:

```

factor(c("SMA", "S1", "SMP", "SD",
        "SD", "S1", "S2"),
       levels = c("SD", "SMP", "SMA", "S1", "S2")) |>
ordered()

```

```

[1] SMA S1  SMP SD  SD  S1  S2
Levels: SD < SMP < SMA < S1 < S2

```

Objek ini disebut *ordered factor*. Skala ordinal dapat diurutkan, tetapi tidak dapat direpresentasikan dengan angka (masih **kualitatif**).

### 1.0.3 Skala Interval

Kita mulai memasuki skala numerik. Artinya, variabel tersebut dapat direpresentasikan **angka**, atau **kuantitatif**. Skala interval memiliki jarak, tetapi tidak ada nol yang berarti. Skala temperatur Celsius dan Fahrenheit tidak memiliki nol. Oleh karena itu, tidak bisa diambil rasio. Atau, jarak antara 16 Agustus dan 17 Agustus adalah satu hari, tapi tidak logis untuk mengatakan satu hari adalah  $n$  kali hari lainnya.

### 1.0.4 Skala Rasio

Memiliki nol yang berarti, sehingga rasio dapat dibandingkan. Jenis-jenis variabel ini termasuk tinggi, berat, lebar, dan lain-lain.

### 1.0.5 Exercise: Skala apa saja?



Figure 1.1: Penguins

species	island	bill_length_mm	bill_depth_mm
Adelie :152	Biscoe :168	Min. :32.10	Min. :13.10
Chinstrap: 68	Dream :124	1st Qu.:39.23	1st Qu.:15.60
Gentoo :124	Torgersen: 52	Median :44.45	Median :17.30
		Mean :43.92	Mean :17.15
		3rd Qu.:48.50	3rd Qu.:18.70
		Max. :59.60	Max. :21.50
		NA's :2	NA's :2

flipper_length_mm	body_mass_g	sex	year
Min. :172.0	Min. :2700	female:165	Min. :2007
1st Qu.:190.0	1st Qu.:3550	male :168	1st Qu.:2007
Median :197.0	Median :4050	NA's : 11	Median :2008
Mean :200.9	Mean :4202		Mean :2008
3rd Qu.:213.0	3rd Qu.:4750		3rd Qu.:2009
Max. :231.0	Max. :6300		Max. :2009
NA's :2	NA's :2		

Tabel berasal dari dataset **Palmer Penguins**. Kalau tabel ini?

agegp	alcgp	tobgp	ncases	ncontrols
25-34:15	0-39g/day:23	0-9g/day:24	Min. : 0.000	Min. : 0.000
35-44:15	40-79 :23	10-19 :24	1st Qu.: 0.000	1st Qu.: 1.000
45-54:16	80-119 :21	20-29 :20	Median : 1.000	Median : 4.000
55-64:16	120+ :21	30+ :20	Mean : 2.273	Mean : 8.807
65-74:15			3rd Qu.: 4.000	3rd Qu.:10.000
75+ :11			Max. :17.000	Max. :60.000

- agegp: Kelompok Usia
- alcgp dan tobgp: Konsumsi alkohol dan tembakau
- ncases dan ncontrols: Jumlah kasus dan kontrol untuk kanker esofagus.

## 1.0.6 Aplikasikan: Mau mencari nilai variabel apa?

Misal kamu ingin tahu lebih lanjut tentang:

1. Asisten,
2. Temanmu di kelas praktikum, dan
3. Gedung CCR.

Apa variabel yang ingin kamu ketahui? Skala pengukurannya apa?



## 1.1 Populasi dan Sampel

Dalam tiap studi tersebut, asisten, teman kelas, dan gedung CCR merupakan **populasi** - seperangkat objek dengan karakteristik tertentu yang kita ingin ketahui. Kita mencari **parameter** populasi - suatu kuantitas yang mendeskripsikan suatu aspek populasi.

### 1.1.1 Aplikasikan: tanya asisten

Kamu punya 2 menit - kumpulkan data dari asisten!

Gunakan variabel-variabel yang kamu telah pikirkan di aplikasi sebelumnya. Cari nilainya!

### 1.1.2 Aplikasikan: tanya teman kuliahmu

Kamu punya 3 menit - kumpulkan data dari teman kuliahmu di ruang praktikum!

Gunakan variabel-variabel yang kamu telah pikirkan di aplikasi sebelumnya. Cari nilainya!

## 1.2 Sampling

Apakah kamu berhasil mengumpulkan data dari **semua** temanmu di kelas ini? Bagaimana jika kamu ingin mengumpulkan data semua mahasiswa IPB, semua warga Bogor, semua warga Indonesia? Terkadang, data populasi **tidak bisa dikumpulkan**, atau **membutuhkan waktu dan tenaga yang banyak**. Pencarian data juga dapat **destruktif** - misal ingin mengetahui ketahanan baterai suatu merk jika dipanaskan - tidak mungkin kita panaskan dan rusaki semua baterai yang diproduksi merk tersebut!

### 1.2.1 Aplikasikan: cari sampel

Kamu memiliki waktu 2 menit. Sekarang tanyakan 7 teman kamu saja!

Tiga orang akan maju dan deskripsikan hasilnya.

### 1.2.2 Statistik

Nilai yang ditemukan dari menanyakan 7 teman adalah **statistik** sampel - suatu kuantitas yang mendeskripsikan suatu aspek sampel. Jika kita menunjukkan grafik data, atau menunjukkan rata-rata/median, kita melakukan **statistika deskriptif**. Namun, Apa yang kita bisa katakan mengenai populasi, dengan data sampel?

Ambil satu variabel yang kamu cari datanya ke 7 teman kelas. Lalu tanyakan ke semua teman; apakah berbeda?

### 1.2.3 Inferensia

Diperlukan **inferensia** untuk menduga parameter populasi dari statistik sampel. Ilmu peluang digunakan untuk menduga dan menentukan ketepatan dugaan tersebut.

## 1.3 Pengumpulan Data

Dari mana asal dataset **Palmer Penguins** dan **Esophageal Cancer**? Pihak lain telah mengumpulkan datanya - ini disebut **data sekunder**. Dalam mata kuliah ini, kita belajar mengumpulkan data, seperti tadi - data yang dikumpulkan sendiri disebut **data primer**.

Bagaimana kita mengumpulkan data untuk mendeskripsikan gedung CCR? Kita telah melakukan **survei** - berinteraksi dengan objek sampel untuk mengumpulkan data. Juga ada **sensus** - mengumpulkan data dari seluruh populasi. Tentu kita tidak dapat menanyakan gedung CCR! Hanya dapat **diobservasi** saja, tanpa interaksi.

Bagaimana jika datanya belum ada? Apa efek mendengarkan asisten sambil jungkir balik pada nilai kuliah? Apakah ada yang kuliah sambil jungkir balik? Harus dilakukan **percobaan** - memberikan perlakuan tertentu, untuk membangkitkan data tersebut.

## 1.4 Exercise

<https://openintro-ims.netlify.app/data-design.html#chp2-exercises>

Kerjakan no 1, 2, 7, 8.

## 2 Konsep Dasar Survey Sampling

Dalam **sensus**, peneliti mengamati seluruh anggota populasi. Karena tenaga dan waktu yang dibutuhkan lama, sensus dilakukan dalam jangka waktu panjang, misal 10 tahun sekali. Sedangkan, di **survei** sebagian anggota populasi diamati.

### 2.1 Terminologi

#### 2.1.1 Elemen

Proses penjelasan terminologi dimulai dari membayangkan pelaksanaan survei. **Siapa yang ingin ada teliti?**

Objek apa yang Anda ingin ketahui karakteristiknya, atau Anda ingin ukur?

Ini adalah elemen.

unsur  $\longrightarrow$  populasi

Grafik menggambarkan agregasi unit dari satu elemen ke kumpulan elemen. **Populasi** adalah koleksi unsur yang ingin diduga karakteristiknya. Sebagai contoh:

satu kendaraan  $\longrightarrow$  kendaraan di Bara

satu orang  $\longrightarrow$  warga Bogor

Warga Bogor adalah koleksi orang di Bogor. Kendaraan di Babakan Raya adalah koleksi kendaraan.

### 2.1.2 Satuan penarikan contoh

Koleksi elemen dari populasi yang tidak **tumpang tindih** dan mencakup seluruh populasi. Maksudnya? Paling sederhana **elemen = satuan penarikan contoh**. Dalam contoh orang, tentu tidak mungkin tumpang tindih. Apakah orang, jika digabung, mencakup seluruh warga Bogor?

Namun, satuan penarikan contoh bisa juga tak sama dengan elemen:

elemen  $\longrightarrow$  sampling unit  $\longrightarrow$  populasi

Sama seperti diagram sebelumnya, panah menunjukkan arah dari unit yang kecil ke unit yang besar. Elemen dapat digabung menjadi sebuah sampling unit. Gabungan sampling unit menjadi populasi. Sebagai contoh:

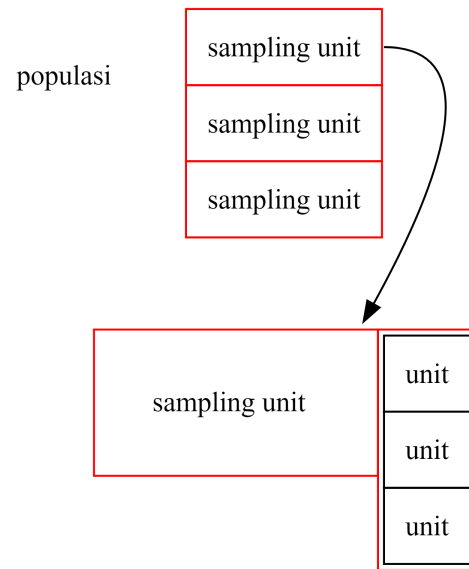
satu mobil  $\longrightarrow$  mobil di menit x  $\longrightarrow$  mobil di Bara

satu orang  $\longrightarrow$  keluarga  $\longrightarrow$  warga Bogor

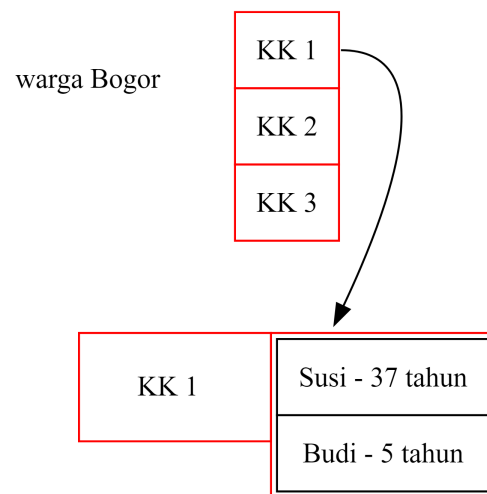
Orang/mobil jika digabung dapat menjadi satu keluarga/satu mobil di interval waktu tertentu. Jika sampling unit digabung, maka menjadi populasi. Note: dalam contoh mobil, ada kemungkinan **tumpang tindih**! Walaupun tiap orang hanya memiliki satu keluarga, bisa jadi mobil yang di menit sebelumnya berada di Babakan Raya ada juga di menit selanjutnya, karena macet!

### 2.1.3 Kerangka

Kerangka adalah **daftar** sampling unit di populasi. Ini dapat digambarkan sebagai suatu tabel, di mana populasi dibagi menjadi beberapa sampling unit. Lalu tiap sampling unit memiliki elemen:



Sebagai contoh:



Pemerintah Kota Bogor memiliki list Kepala Keluarga; lalu, di tiap KK tersebut ada suatu keluarga yang beranggotakan elemen bernama Susi dan Budi.

### 2.1.4 Sampel

Sampel adalah kumpulan satuan penarikan contoh dari kerangka. Misal, ambil KK 1 dan 2. Hitung mobil di jam-jam tertentu

### 2.1.5 Exercise

1. Cari suatu unit!
2. Kumpulan unit seperti apa yang kamu ingin duga karakteristiknya!
3. Kumpulan unit yang lebih kecil dari populasi apa saja yang mungkin terjadi?
4. Bagaimana bentuk kerangkanya?
5. Beri contoh pengambilan sampel dari kerangka tersebut.

## 2.2 Sampling

### 2.2.1 Probability sampling

Dalam probability sampling, tiap anggota populasi memiliki probabilitas untuk dipilih; probabilitasnya tidak perlu sama. Bisa saja:

1. Probabilitas semua anggota dipilih sama.
2. Anggota kelompok tertentu, yang memang memiliki proporsi lebih besar di kerangka, memiliki probabilitas dipilih lebih besar.
3. Cara lain.

### 2.2.2 Non-probability sampling

Probabilitas anggota populasi dipilih tidak diketahui di non-probability sampling.

### 2.2.3 Sampling error

Berapa banyak dari kalian yang suka matematika?

Apa kesimpulan yang dapat diambil mengenai minat matematika mahasiswa IPB? Kemungkinan susah untuk mengambil kesimpulan; mungkin saja mahasiswa Statistika memiliki kemampuan matematika lebih tinggi, atau ketertarikan matematika lebih tinggi. Mungkin saja mahasiswa Statistika terlalu sering terekspos matematika sehingga tidak menyukai pelajaran tersebut. Jika kita mengamati sebagian populasi (contoh), akan ada kesalahan karena

contoh tersebut belum tentu mewakili keragaman populasi. Jika mahasiswa dari berbagai jurusan diamati, kesimpulan lebih valid. Lalu, dengan **probability sampling** tingkat kesalahan dapat diduga.

## 2.3 Non-sampling errors

### 2.3.1 Nonobservation

Kesalahan yang terjadi karena gagal mengobservasi elemen:

#### 2.3.1.1 *Non-coverage*

Jika kerangka sampling tidak mencakup seluruh populasi, misal DPT tidak lengkap atau tidak semua pengemudi punya SIM!

#### 2.3.1.2 *Non-response*

Error ini sering lebih fatal. Beberapa jenis error:

1. Tidak bisa mengontak unit

Tanpa keluar CCR, misal Anda coba survei jumlah pekerja dan pemasok tempat makan kalian hari ini. Anda mungkin tak memiliki kontak elemen survei tersebut! Namun, jika Anda menanyakan apa yang dimakan teman Anda hari ini, hal tersebut mungkin ditemukan.

Misal, dengan contoh acak sederhana tiap elemen memiliki probabilitas  $1/n$  untuk disampel. Apakah ada bisa menghitung probabilitas elemen disampel jika sebenarnya dipilih elemen lain, elemen lain tidak ada, lalu diganti elemen yang dekat? Ini susah.

Lalu, misal Anda melakukan survey ke rumah warga di jam siang. Ternyata, orang berusia dewasa sedang kerja. Apa yang terjadi jika yang ditanya adalah orang yang ada di rumah?

- Siapa yang mungkin di rumah warga pada jam siang?
- Apakah mungkin profil orang tersebut beda dengan warga yang Anda cari?

2. Unit tak bisa menjawab pertanyaan

Buat pertanyaan yang susah dijawab orang. Bukan pertanyaan yang sensitif, tetapi susah dimengerti/susah dicari jawabannya. Misal, apakah Anda mengingat informasi:

1. Pengeluaran dalam minggu ini.
2. Jarak berjalan kaki dalam hari ini.
3. Jumlah teman yang dikontak melalui WhatsApp dua hari ini.

Belum tentu responden mengetahui/mengingat informasi yang kita ingin tanyakan.

3. Unit tidak ingin menjawab pertanyaan

Ini cukup jelas.

4. Dishonest interviewer

Selain kesalahan dari responden, interviewer dapat tidak jujur dan mengisi survei dengan sendirinya.

### **2.3.2 Errors of observation**

Informasi dari elemen dapat diobservasi, tetapi diobservasi dengan salah.

#### **2.3.2.1 Interviewer yang tidak netral**

Apakah kalian suka cari masalah dengan orang? Jika interviewer dirasa mendukung posisi tertentu, responden mungkin saja mencoba mengikuti posisi interviewer, atau melawan posisi tersebut.

#### **2.3.2.2 Kesalahan responden**

1. Jika Anda ditanya harga outfit Anda (dan kebetulan bagus outfitnya), apakah Anda ingin menaikkan harga outfit tersebut ke interviewer?
2. Jika Anda ditanya pernah melanggar aturan kampus, apakah Anda akan menjawab ya atau tidak?
3. Apakah Anda akan mengatakan jumlah uang bulanan Anda secara tepat jika ditanya?

Inti dari aktivitas ini: responden memiliki motivasi yang berbeda-beda untuk menjawab pertanyaan. Misal, responden ingin tampak mengikuti aturan, ingin tampak kaya (atau menyembunyikan kekayaan), dan lain-lain.

#### **2.3.2.3 Instrumen**

Apa definisi Anda untuk:

1. Anak
2. Pengangguran
3. Harga makanan murah

Apakah definisi semua orang sama? Bagaimana jika survei menanyakan suatu hal yang definisinya tidak jelas?



#### 2.3.2.4 Input Data

Misal data sudah dikumpulkan. Apakah Anda pernah salah ketik? Apakah komputer Anda pernah mengalami *bug*? Data tersebut mungkin akan berubah.

## 2.4 Mengurangi kesalahan

### 2.4.1 Trade-off: sampling error vs non-sampling

Mana yang lebih mengikuti kaidah percontohan: mengambil contoh acak mahasiswa IPB atau menanya teman? Siapa yang lebih mungkin merespon: teman atau mahasiswa yang dicontoh acak?

Semakin besar ukuran sampel, semakin mendekati populasi. Kira-kira, apakah Anda akan lebih lelah menanyakan lebih banyak orang lalu memasukkan datanya? Jika lebih lelah, apakah Anda lebih mungkin salah?

### 2.4.2 Cara mengurangi non-sampling error

#### 1. Callback

Apakah chat Anda pernah tidak direspon? Apa yang Anda lakukan? Jika di-chat lagi, apakah biasanya ada beberapa yang merespon kembali? Sama saja, di survei, surveyor dapat kembali ke responden yang belum merespon dan menanyakan pertanyaan yang sama lagi.

#### 2. Rewards

Jika mengisi, dapat uang/dst. Pertanyaannya: apakah orang yang ingin mengisi survei untuk dapat *reward* beda dengan orang pada umumnya? Mungkin orang tersebut memiliki keadaan finansial tertentu, atau lebih suka uang.

#### 3. Melatih interviewer

Interview dapat dilatih agar tampak netral dan mengetahui jawaban jujur/tidak jujur

#### 4. Data checking

Apakah mungkin:

- Seseorang berusia 1000 tahun?
- Satu keluarga memiliki 50 anak?
- Orang dewasa yang sudah menikah berusia 5 tahun?

Data-data tersebut perlu diperiksa.

## 5. Memperbaiki kuesioner

Beberapa aspek yang dapat diperhatikan:

- Urutan pertanyaan

Biasanya, responden ingin konsisten dengan jawabannya. Misal, ditanyakan “Apakah Anda setuju dengan pemotongan subsidi BBM untuk peningkatan dana pendidikan?”. Jika setelah pertanyaan itu ditanyakan “Apakah Anda setuju dengan pemotongan subsidi BBM?”, responden lebih mungkin menjawab “Ya”. Urutan tersebut perlu diperhatikan.

- Pertanyaan terbuka vs. tertutup

Pertanyaan terbuka memberi lebih banyak opsi bagi responden, tapi susah diproses (non-sampling error meningkat). Pertanyaan tertutup dapat lebih mudah diproses, tetapi harus dipastikan semua opsi yang hendak dipilih responden ada.

- Opsi

Misal, opsi netral. Kebanyakan responden mungkin memilih opsi netral. Apakah lebih baik untuk memaksa responden memilih?

- Wording

Responden mungkin tidak menyukai kata yang sangat negatif seperti “melarang”, tetapi jika responden ditanya “Apakah setuju dengan membolehkan X”, bisa jadi banyak responden menjawab “Tidak”.

## 2.5 Cara pengumpulan data

### 2.5.1 Personal interview:

- Apakah Anda akan merespon orang yang menanyakan Anda secara langsung?
- Biasanya, Anda bicara mengikuti prosedur tertentu, atau mengikuti saja alur perancangannya bagaimana?

Personal interview biasanya memiliki tingkat respon tinggi, tetapi bias mungkin muncul dari interviewer.

### 2.5.2 Telephone interview

- Apakah semua kontak di HP Anda nomor orang? Apakah ada nomor bisnis? Nomor penipu? Nomor lama yang sudah tidak dipakai lagi? Mungkin saja kerangka percontohan tak akurat.
- Biasanya, berapa lama Anda berbicara di telepon? Orang biasanya tak ingin berbicara lama.

### 2.5.3 Self-Administered Questionnaire

- Tidak bisa memastikan responden mengisi atau tidak. Jika tidak ada yang mengingatkan/mengawasi, apakah Anda pernah lupa mengisi form?

Cara ini sangat mudah, tetapi response rate rawan rendah.

### 2.5.4 Observasi

Misal Anda ingin meneliti berapa sering seseorang masuk kuliah. Apakah Anda akan:

1. Melihat absensi?
2. Menanyakan orang tersebut?

Kadang, beberapa data lebih baik diobservasi langsung.

## 2.6 Tahapan pembuatan survey

Lakukan tahap 1-3!

1. Tujuan survei apa?
2. Populasi target apa?
3. Dari mana Anda memperoleh sampling frame?

Lalu lakukan tahap 5-7!

5. Apakah Anda akan mewawancarai responden?
6. Bagaimana cara Anda memastikan kuesioner reliabel dan valid?
7. Apakah pewawancara perlu dilatih; seperti apa?

Terakhir, lakukan tahap 9-12!

9. Bagaimana Anda mengumpulkan data dari responden?
10. Data ditaruh di mana?

11. Bagaimana Anda menganalisis data?
12. Bagaimana Anda melaporkan hasil survei?

## 3 Simple Random Sampling

Survei adalah **menduga** parameter populasi berdasarsarkan informasi sampel. Faktor apa saja yang memengaruhi jumlah informasi di sampel?

1. Ukuran sampel
2. Keragaman di sampel - dapat dikontrol dengan rancangan percontohan yang baik. Paling sederhana: **simple random sampling**

Dua poin penting:

1. Semua elemen memiliki **peluang sama** untuk dipilih.
2. Pemilihan elemen **saling bebas**. Artinya, ada atau tidaknya elemen di sampel tidak memengaruhi probabilitas elemen lain dipilih.

### 3.0.1 Exercise: apakah ini sampel acak sederhana?

Do these methods produce a simple random sample of students from a class of 30 students?

1. Select the first six students on the class roll sheet (*absensi*).
2. Pick a digit at random and select those students whose phone numbers end in that digit.
3. If the classroom has six rows of chairs with five seats in each row, choose a row at random and select all students in that row.
4. If the class consists of 15 boys and 15 girls, assign the boys the numbers from 1 to 15, and the girls the numbers from 16 to 30. Then use a random digit table to select six numbers from 1 to 30. Select the students assigned those numbers in your sample.
5. If the class consists of 15 boys and 15 girls, assign the boys the numbers from 1 to 15, and the girls the numbers from 16 to 30. Then use a random digit table to select three numbers from 1 to 15 and three numbers from 16 to 30. Select the students assigned those numbers in your sample.
6. Randomly choose a letter from the English alphabet and select for the sample those students whose last names begin with that letter. If no last name begins with that letter, randomly choose another letter from the alphabet.

Kunci dari jawaban tersebut adalah:

1. Select the first six students on the class roll sheet (*absensi*) - **peluang tak sama**. Enam orang pertama pasti terpilih, orang lainnya tidak mungkin terpilih.

2. Pick a digit at random and select those students whose phone numbers end in that digit - **sampel acak sederhana**. Alokasi nomor telpon cukup acak dan pemilihan juga acak.
3. If the classroom has six rows of chairs with five seats in each row, choose a row at random and select all students in that row - **tidak saling bebas**. Jika elemen di baris tersebut terpilih, pasti elemen lain juga terpilih. Lalu, elemen di baris lain pasti tidak terpilih.
4. If the class consists of 15 boys and 15 girls, assign the boys the numbers from 1 to 15, and the girls the numbers from 16 to 30. Then use a random digit table to select six numbers from 1 to 30. Select the students assigned those numbers in your sample - **sampel acak sederhana**.
5. If the class consists of 15 boys and 15 girls, assign the boys the numbers from 1 to 15, and the girls the numbers from 16 to 30. Then use a random digit table to select three numbers from 1 to 15 and three numbers from 16 to 30. Select the students assigned those numbers in your sample - **tidak saling bebas**. Jika 3 laki-laki dipilih, pasti tidak mungkin laki-laki lain dipilih.
6. Randomly choose a letter from the English alphabet and select for the sample those students whose last names begin with that letter. If no last name begins with that letter, randomly choose another letter from the alphabet. **Apakah nama-nama yang diawali alfabet tertentu lebih mungkin daripada alfabet lain?** Selain itu, mungkin tidak saling bebas karena jika nama akhir yang banyak dimiliki orang terpilih; orang dengan nama lain tak mungkin terpilih.

### 3.0.2 Contoh - Rating TV

- Apakah Anda memiliki acara televisi favorit?
- Apakah acara favorit Anda pernah mengalami berhenti tayang (*cancellation*)?
- Bagaimana stasiun televisi menentukan acara mana yang diberhentikan?

Nielsen Ratings - ambil 5000 rumah tangga di AS; tidak boleh ada relawan. Dipasang meter elektronik yang mengetahui acara yang sedang ditonton.

### 3.0.3 Cara Mengambil Sampel Acak Sederhana

Cara salah:

- Haphazard: sesuai keinginan peneliti
- Representative: ambil sampel yang dianggap mewakili populasi

Peneliti bisa saja berbias dan walaupun representatif, **tingkat kesalahan tak dapat diketahui** karena tidak diketahui struktur probabilitasnya.

Cara benar:

- Undian - masukkan angka 1 sampai n, ambil!

- Tabel angka acak - biasanya menggunakan komputer!

### 3.0.4 Exercise: deskripsikan bias

1. A student wants to determine the average size of farms in a county in Iowa. He drops some rice randomly on a map of the county and uses the farms hit by grains of rice as the sample (*county - kabupaten*).
2. To find the average length of string in a bag, a student reaches in, mixes up the strings, selects one, mixes them up again, selects another, and so on (*string - tali*).
3. To estimate the percentage of students who passed the first Advanced Placement Statistics exam, a teacher on an Internet discussion list for teachers of AP Statistics asked teachers on the list to report to him how many of their students took the test and how many passed (*discussion list - forum diskusi*).
4. In 1984, Ann Landers conducted a poll on the marital happiness of women by asking women to write to her (*marital - pernikahan*).
5. In a study about whether valedictorians “succeed big in life,” a professor “traveled across Illinois, attending high school graduations and selecting 81 students to participate. . . . He picked students from the most diverse communities possible, from little rural schools to rich suburban schools near Chicago to city schools.”

*Valedictorian: lulusan terbaik.*

Jawaban dari exercise tersebut adalah:

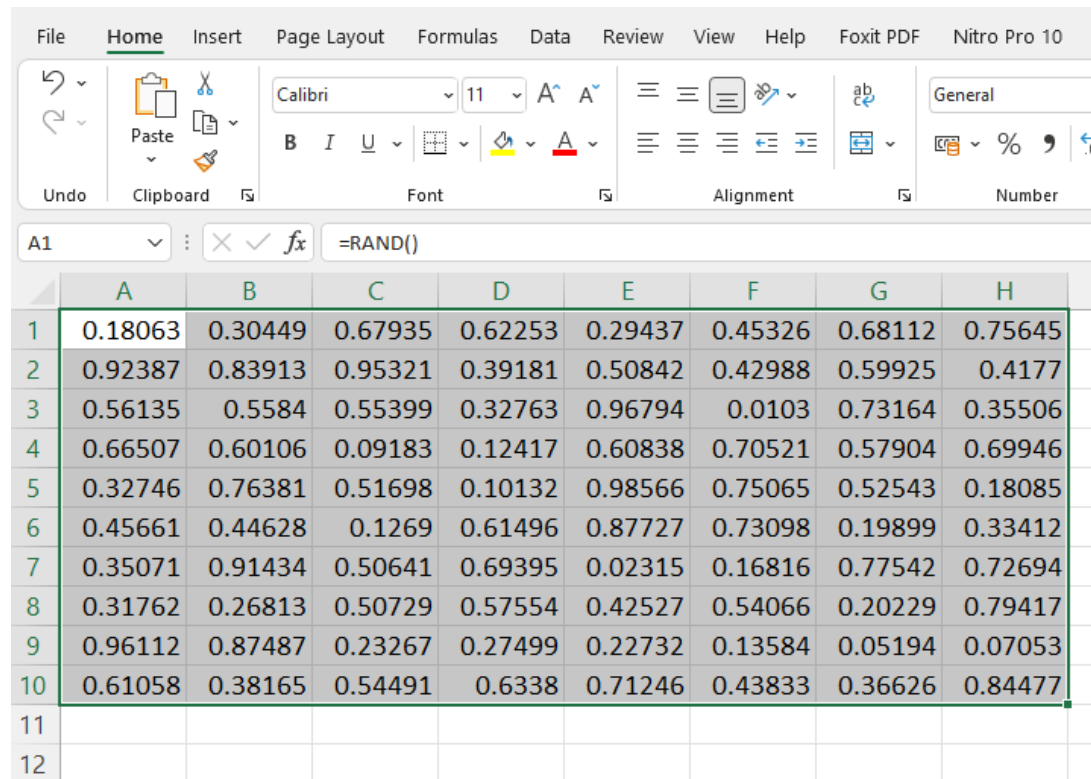
1. A student wants to determine the average size of farms in a county in Iowa. He drops some rice randomly on a map of the county and uses the farms hit by grains of rice as the sample (*county - kabupaten*) - **proses ini cukup acak.**
2. To find the average length of string in a bag, a student reaches in, mixes up the strings, selects one, mixes them up again, selects another, and so on (*string - tali*) - **proses ini juga cukup acak.**
3. To estimate the percentage of students who passed the first Advanced Placement Statistics exam, a teacher on an Internet discussion list for teachers of AP Statistics asked teachers on the list to report to him how many of their students took the test and how many passed (*discussion list - forum diskusi*) - **ada bias yang muncul karena guru-guru yang menjawab di forum diskusi belum tentu representatif terhadap guru umumnya.** Mungkin, guru tersebut lebih ambisius.
4. In 1984, Ann Landers conducted a poll on the marital happiness of women by asking women to write to her (*marital - pernikahan*). **Bias yang sama.** Orang yang menulis tentang pernikahannya mungkin memiliki perasaan ekstrim, seperti sangat senang atau sangat marah.
5. In a study about whether valedictorians “succeed big in life,” a professor “traveled across Illinois, attending high school graduations and selecting 81 students to participate. . . .

He picked students from the most diverse communities possible, from little rural schools to rich suburban schools near Chicago to city schools.”

**Walaupun representatif, tingkat kesalahan tak dapat diperkirakan.**

Untuk melakukan pengacakan, sering dipakai *random number generator*:

#### 3.0.4.1 Excel



The screenshot shows the Microsoft Excel interface with the 'Home' tab selected. The formula bar at the top displays '=RAND()'. The worksheet grid shows columns A through H and rows 1 through 12. Each cell in the grid (A1:H10) contains a random decimal number generated by the RAND() function. The numbers are distributed across the range 0 to 1. The interface includes standard Excel menus (File, Home, Insert, etc.) and toolbars for font, alignment, and numbers.

	A	B	C	D	E	F	G	H
1	0.18063	0.30449	0.67935	0.62253	0.29437	0.45326	0.68112	0.75645
2	0.92387	0.83913	0.95321	0.39181	0.50842	0.42988	0.59925	0.4177
3	0.56135	0.5584	0.55399	0.32763	0.96794	0.0103	0.73164	0.35506
4	0.66507	0.60106	0.09183	0.12417	0.60838	0.70521	0.57904	0.69946
5	0.32746	0.76381	0.51698	0.10132	0.98566	0.75065	0.52543	0.18085
6	0.45661	0.44628	0.1269	0.61496	0.87727	0.73098	0.19899	0.33412
7	0.35071	0.91434	0.50641	0.69395	0.02315	0.16816	0.77542	0.72694
8	0.31762	0.26813	0.50729	0.57554	0.42527	0.54066	0.20229	0.79417
9	0.96112	0.87487	0.23267	0.27499	0.22732	0.13584	0.05194	0.07053
10	0.61058	0.38165	0.54491	0.6338	0.71246	0.43833	0.36626	0.84477
11								
12								

Gunakan fungsi =RAND():

#### 3.0.4.2 R

Gunakan fungsi `runif()`:

```
runif(100) |> head(5)
```

```
[1] 0.1616881 0.2276692 0.7252934 0.4352142 0.3618568
```



### 3.0.4.3 Python

Gunakan rng dari package `numpy`:

```
import numpy as np

rng = np.random.default_rng(3854)
rng.random(5)
```

```
array([0.50955422, 0.53163209, 0.98811285, 0.47079458, 0.08600762])
```

## 3.1 Memilih sampel acak sederhana

Algoritma:

1. Buat angka acak sejumlah elemen yang ada
2. Pasangkan tiap elemen dengan angka acak
3. Urutkan elemen sesuai angka acak, pilih  $n$  teratas.

### 3.1.1 Step 0: Lihat Dataset dan Jumlah Row

Sebelum melaksanakan algoritma tersebut, lihat dataset dan jumlah row:

#### 3.1.1.1 Excel

Pertama, ambil data dari csv:

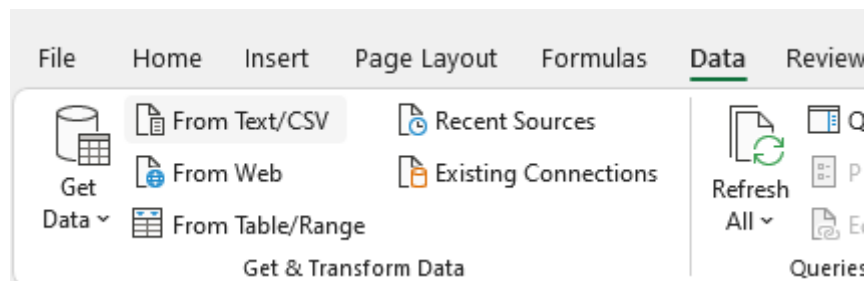


Figure 3.1: Pengambilan data

Pilih file, lalu preview dan load:

Data akan terlihat di Excel.

penguins.csv

File Origin: 1252: Western European (Windows) Delimiter: Comma Data Type Detection: Based on first 200 rows

rowid	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
3	Adelie	Torgersen	40.3	18	195	3250	female	2007
4	Adelie	Torgersen	NA	NA	NA	NA	NA	2007
5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
6	Adelie	Torgersen	39.3	20.6	190	3650	male	2007
7	Adelie	Torgersen	38.9	17.8	181	3625	female	2007
8	Adelie	Torgersen	39.2	19.6	195	4675	male	2007
9	Adelie	Torgersen	34.1	18.1	193	3475	NA	2007
10	Adelie	Torgersen	42	20.2	190	4250	NA	2007
11	Adelie	Torgersen	37.8	17.1	186	3300	NA	2007
12	Adelie	Torgersen	37.8	17.3	180	3700	NA	2007
13	Adelie	Torgersen	41.1	17.6	182	3200	female	2007
14	Adelie	Torgersen	38.6	21.2	191	3800	male	2007
15	Adelie	Torgersen	34.6	21.1	198	4400	male	2007
16	Adelie	Torgersen	36.6	17.8	185	3700	female	2007
17	Adelie	Torgersen	38.7	19	195	3450	female	2007
18	Adelie	Torgersen	42.5	20.7	197	4500	male	2007
19	Adelie	Torgersen	34.4	18.4	184	3325	female	2007
20	Adelie	Torgersen	46	21.5	194	4200	male	2007

The data in the preview has been truncated due to size limits.

Load Transform Data Cancel

Figure 3.2: Preview

### 3.1.1.2 R

Langkah relatif sama. Baca CSV, lihat file.

```
penguins <- read.csv("penguins.csv")

penguins |> head(3)
```

	rowid	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm
1	1	Adelie	Torgersen	39.1	18.7	181
2	2	Adelie	Torgersen	39.5	17.4	186
3	3	Adelie	Torgersen	40.3	18.0	195
		body_mass_g	sex	year		
1		3750	male	2007		
2		3800	female	2007		
3		3250	female	2007		

```
penguins |> nrow()
```

```
[1] 344
```

### 3.1.1.3 Python

```
import pandas as pd

penguins = pd.read_csv("penguins.csv")
print(penguins.head())
```

	rowid	species	island	...	body_mass_g	sex	year
0	1	Adelie	Torgersen	...	3750.0	male	2007
1	2	Adelie	Torgersen	...	3800.0	female	2007
2	3	Adelie	Torgersen	...	3250.0	female	2007
3	4	Adelie	Torgersen	...	NaN	NaN	2007
4	5	Adelie	Torgersen	...	3450.0	female	2007

```
[5 rows x 9 columns]
```

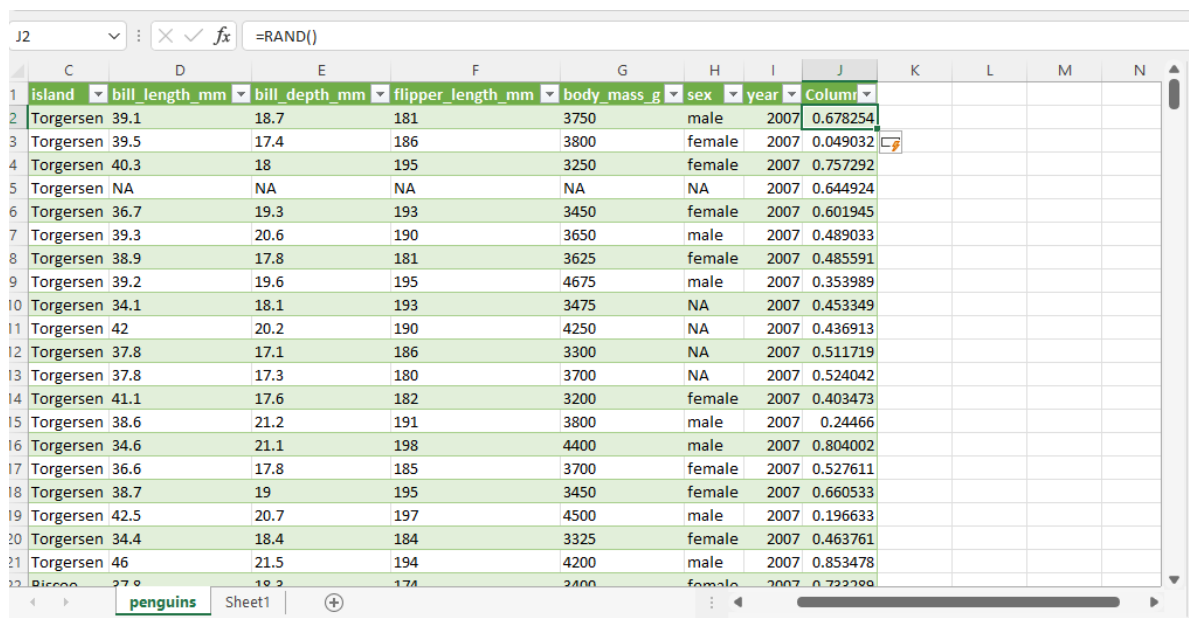
```
print(len(penguins.index))
```

### 3.1.2 Step 1-2: Angka Acak, Pasangan

Lalu, buat angka acak. Angka acak ini merupakan kolom baru di dataset. Tiap elemen mendapat angka acak yang unik. Bagaimana kita tahu jumlah angka acak yang perlu dibuat? Cari terlebih dahulu jumlah elemen di dataset.

#### 3.1.2.1 Excel

Buat angka acak dengan `rand()`.



	C	D	E	F	G	H	I	J	K	L	M	N
1	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year	Column				
2	Torgersen	39.1	18.7	181	3750	male	2007	0.678254				
3	Torgersen	39.5	17.4	186	3800	female	2007	0.049032				
4	Torgersen	40.3	18	195	3250	female	2007	0.757292				
5	Torgersen	NA	NA	NA	NA	NA	2007	0.644924				
6	Torgersen	36.7	19.3	193	3450	female	2007	0.601945				
7	Torgersen	39.3	20.6	190	3650	male	2007	0.489033				
8	Torgersen	38.9	17.8	181	3625	female	2007	0.485591				
9	Torgersen	39.2	19.6	195	4675	male	2007	0.353989				
10	Torgersen	34.1	18.1	193	3475	NA	2007	0.453349				
11	Torgersen	42	20.2	190	4250	NA	2007	0.436913				
12	Torgersen	37.8	17.1	186	3300	NA	2007	0.511719				
13	Torgersen	37.8	17.3	180	3700	NA	2007	0.524042				
14	Torgersen	41.1	17.6	182	3200	female	2007	0.403473				
15	Torgersen	38.6	21.2	191	3800	male	2007	0.24466				
16	Torgersen	34.6	21.1	198	4400	male	2007	0.804002				
17	Torgersen	36.6	17.8	185	3700	female	2007	0.527611				
18	Torgersen	38.7	19	195	3450	female	2007	0.660533				
19	Torgersen	42.5	20.7	197	4500	male	2007	0.196633				
20	Torgersen	34.4	18.4	184	3325	female	2007	0.463761				
21	Torgersen	46	21.5	194	4200	male	2007	0.853478				
22	Biscoe	37.9	19.2	174	3400	female	2007	0.722389				

Figure 3.3: Angka acak

Namun, angka acak ini akan selalu berubah jika di-*sort*. Oleh karena itu, *copy*, (**CTRL+C**) lalu paste *as value*. Opsi *paste as value* ditemukan dengan meng-klik kanan:

Lalu ditemukan opsi tersebut; opsi berupa suatu *clipboard* (kertas di atas papan jalan) dengan angka 123:

#### 3.1.2.2 R

Jumlah angka acak dicari menggunakan `nrow` dari dataset. `Mutate` menghasilkan peubah baru.

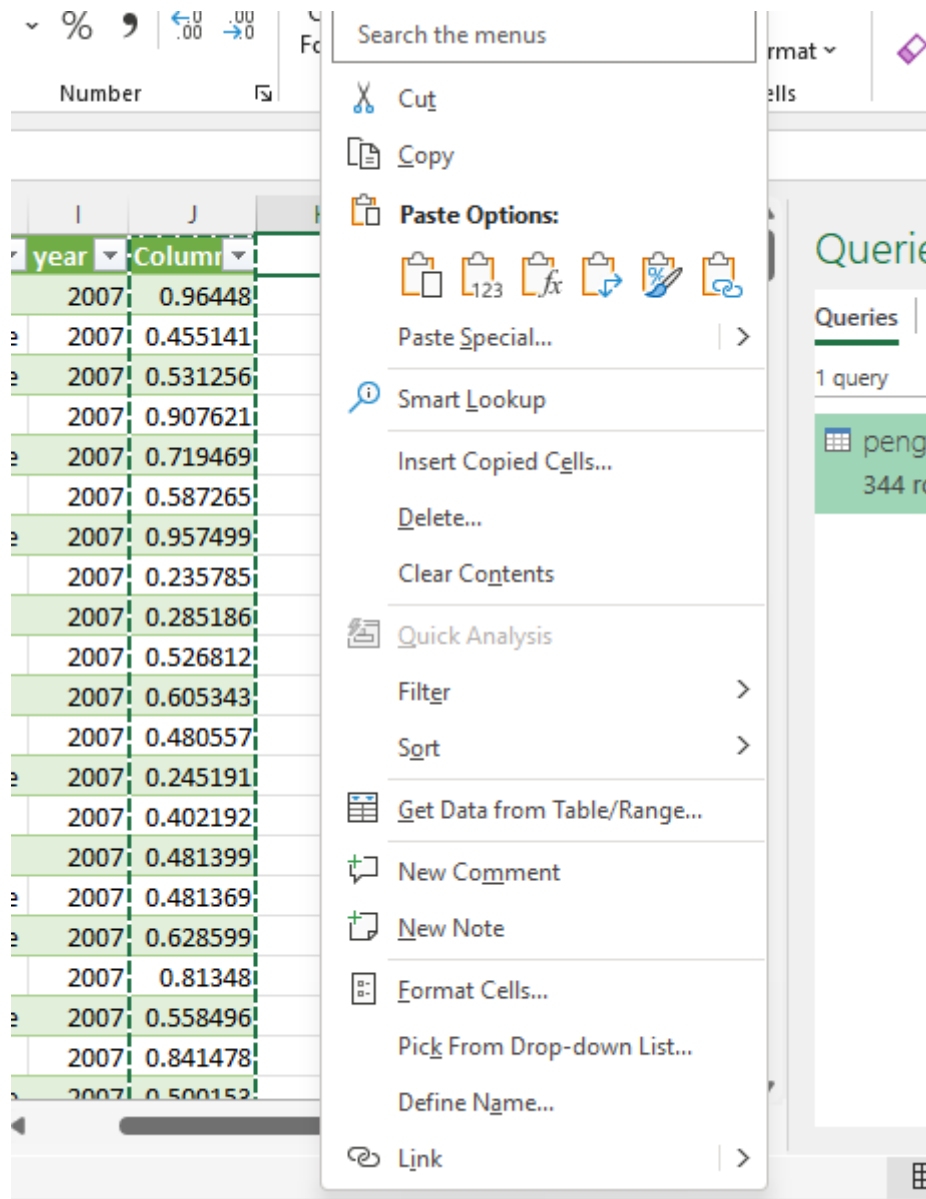


Figure 3.4: Menu paste



```
library(dplyr)
```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

filter, lag

The following objects are masked from 'package:base':

intersect, setdiff, setequal, union

```
penguins <- read.csv("penguins.csv")
penguins <- penguins |> mutate(rando = runif(nrow(penguins)))

penguins |> head(3) |> knitr::kable()
```

rowid	species	island	bill_length	bill_depth	flipper_length	body_mass	sex	year	rando
1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007	0.1827843
2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007	0.2001530
3	Adelie	Torgersen	40.3	18.0	195	3250	female	2007	0.6075793

### 3.1.2.3 Python

Algoritma sama. `namaDataset.insert` digunakan untuk memasukkan angka acak.

```
import pandas as pd
import numpy as np
penguins = pd.read_csv("penguins.csv") #load csv

#generate random number
rng = np.random.default_rng(3854)
rando = rng.random(len(penguins.index))

penguins.insert(loc = 0, column = 'randomNumber', value = rando) #insert random number
penguins.head() #show
```

	randomNumber	rowid	species	...	body_mass_g	sex	year
0	0.509554	1	Adelie	...	3750.0	male	2007
1	0.531632	2	Adelie	...	3800.0	female	2007
2	0.988113	3	Adelie	...	3250.0	female	2007
3	0.470795	4	Adelie	...	NaN	NaN	2007
4	0.086008	5	Adelie	...	3450.0	female	2007

[5 rows x 10 columns]

### 3.1.3 Step 3: Sort, Ambil

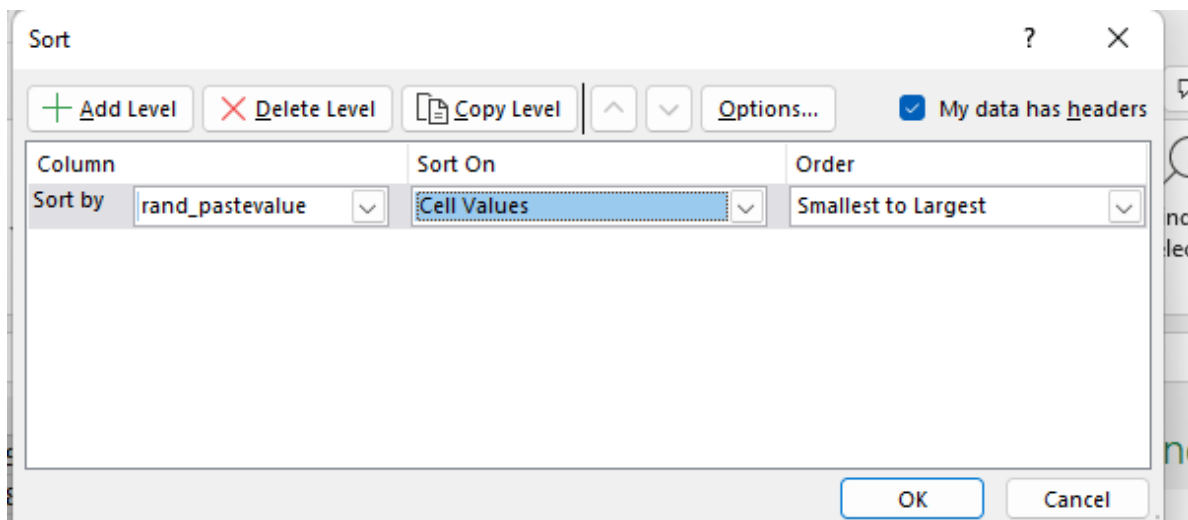
Lalu, sortir data-nya dan ambil n data teratas:

#### 3.1.3.1 Excel

Jika sudah berbentuk tabel, klik kolom nilai acak lalu *sort* sesuai keinginan:

Jika belum berbentuk tabel, pilih *sampling frame* yang ingin disortir, lalu klik *sort & filter*.

Lalu, pilih kolom nilai acak dan sortir.



#### 3.1.3.2 R

Sortir menggunakan *arrange*.



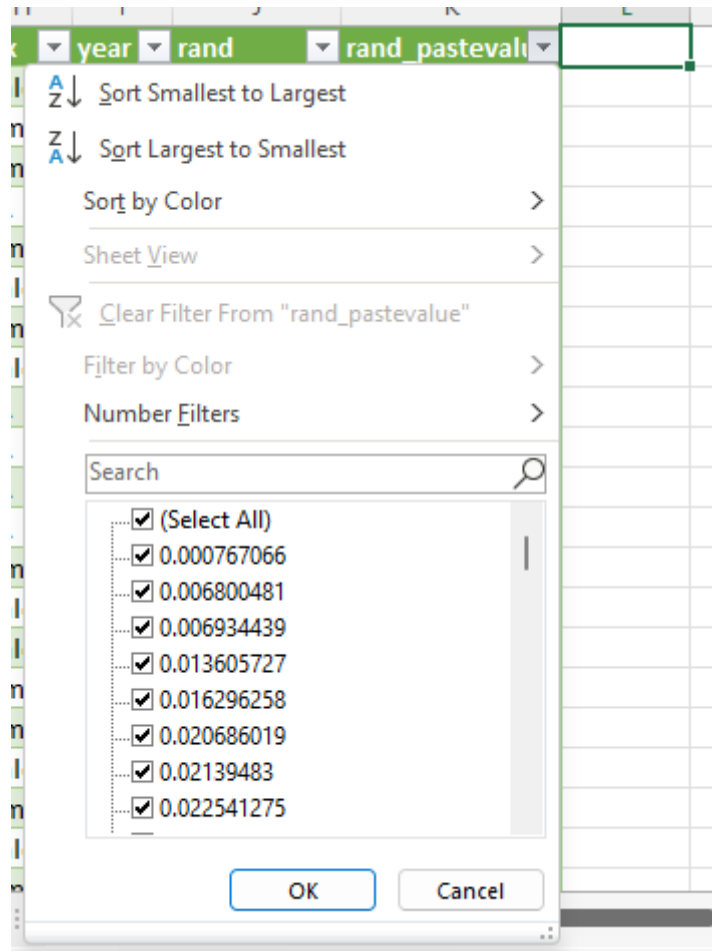


Figure 3.6: Sortir, jika tabel

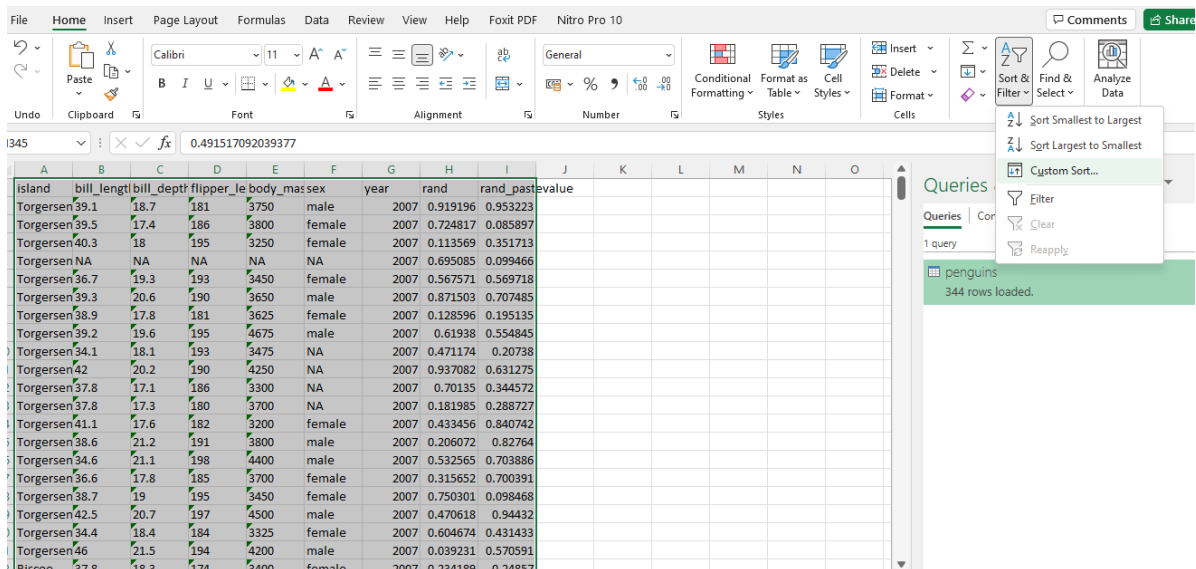


Figure 3.7: Sortir, tanpa tabel

```
penguins |> arrange(desc(rando)) |>
  head(3) |> knitr::kable()
```

rowid	species	island	bill_length	bill_depth	flipper_length	body_mass_g	sex	year	rando
88	Adelie	Dream	36.9	18.6	189	3500	female	2008	0.9909071
208	Gentoo	Biscoe	45.0	15.4	220	5050	male	2008	0.9892091
254	Gentoo	Biscoe	55.9	17.0	228	5600	male	2009	0.9871470

### 3.1.3.3 Python

Gunakan `namaDataset.sort_values(by = "kolom angka acak", ...)`

```
import pandas as pd

penguins_sorted = penguins.sort_values(by = "randomNumber", ascending=False)
print(penguins_sorted.head(58))
```

	randomNumber	rowid	species	... body_mass_g	sex	year
82	0.998006	83	Adelie	... 3800.0	female	2008
2	0.988113	3	Adelie	... 3250.0	female	2007
179	0.985227	180	Gentoo	... 5650.0	male	2007

64	0.982089	65	Adelie	...	2850.0	female	2008
20	0.981606	21	Adelie	...	3400.0	female	2007
316	0.979618	317	Chinstrap	...	3950.0	male	2008
36	0.978045	37	Adelie	...	3950.0	male	2007
226	0.971176	227	Gentoo	...	4700.0	female	2008
88	0.968786	89	Adelie	...	3950.0	male	2008
291	0.967539	292	Chinstrap	...	4050.0	male	2007
142	0.964342	143	Adelie	...	3050.0	female	2009
66	0.964280	67	Adelie	...	3350.0	female	2008
192	0.964219	193	Gentoo	...	3950.0	female	2008
245	0.962202	246	Gentoo	...	5650.0	male	2009
337	0.957290	338	Chinstrap	...	3650.0	female	2009
261	0.955021	262	Gentoo	...	5500.0	male	2009
215	0.951514	216	Gentoo	...	5650.0	male	2008
91	0.949608	92	Adelie	...	4300.0	male	2008
73	0.949112	74	Adelie	...	4150.0	male	2008
223	0.938686	224	Gentoo	...	5000.0	male	2008
289	0.929323	290	Chinstrap	...	4050.0	male	2007
242	0.923924	243	Gentoo	...	4950.0	female	2009
317	0.914438	318	Chinstrap	...	3650.0	female	2008
19	0.908108	20	Adelie	...	4200.0	male	2007
197	0.906972	198	Gentoo	...	4900.0	female	2008
177	0.905509	178	Gentoo	...	5100.0	male	2007
229	0.901730	230	Gentoo	...	6000.0	male	2008
54	0.900127	55	Adelie	...	2900.0	female	2008
156	0.898963	157	Gentoo	...	5400.0	male	2007
225	0.891437	226	Gentoo	...	5200.0	female	2008
239	0.887460	240	Gentoo	...	5300.0	male	2009
273	0.884751	274	Gentoo	...	5750.0	male	2009
324	0.883531	325	Chinstrap	...	3250.0	male	2009
23	0.881101	24	Adelie	...	3950.0	male	2007
27	0.880034	28	Adelie	...	3200.0	female	2007
175	0.876310	176	Gentoo	...	5050.0	male	2007
134	0.876252	135	Adelie	...	3425.0	female	2009
310	0.869858	311	Chinstrap	...	3600.0	male	2008
272	0.866883	273	Gentoo	...	4850.0	female	2009
106	0.861813	107	Adelie	...	3750.0	female	2009
11	0.860677	12	Adelie	...	3700.0	NaN	2007
280	0.857667	281	Chinstrap	...	3725.0	male	2007
320	0.857076	321	Chinstrap	...	3675.0	female	2009
70	0.855720	71	Adelie	...	3600.0	female	2008
297	0.853891	298	Chinstrap	...	3400.0	male	2007
189	0.849117	190	Gentoo	...	5250.0	male	2008

110	0.832433	111	Adelie	...	3825.0	female	2009
100	0.829604	101	Adelie	...	3725.0	female	2009
178	0.828691	179	Gentoo	...	4100.0	NaN	2007
234	0.828592	235	Gentoo	...	4725.0	female	2009
194	0.824546	195	Gentoo	...	4300.0	female	2008
195	0.821451	196	Gentoo	...	4750.0	male	2008
202	0.819321	203	Gentoo	...	4850.0	female	2008
247	0.818314	248	Gentoo	...	5200.0	male	2009
307	0.817592	308	Chinstrap	...	4300.0	male	2008
224	0.814083	225	Gentoo	...	5100.0	male	2008
308	0.813684	309	Chinstrap	...	3350.0	female	2008
257	0.807059	258	Gentoo	...	5500.0	male	2009

[58 rows x 10 columns]

### 3.1.4 Alternatif

1. Buat  $n$  integer random dari 1 ke  $N$ , jumlah populasi.
2. Ambil row yang sesuai integer random tersebut.

#### 3.1.4.1 R

```
library(dplyr)

indexes <- sample.int(n = nrow(penguins), size = 3)
penguins |> filter(rowid %in% indexes) |> knitr::kable()
```

rowid	species	island	bill_length	bill_depth	flipper_length	body_mass	sex	year	rando
137	Adelie	Dream	35.6	17.5	191	3175	female	2009	0.6839653
150	Adelie	Dream	37.8	18.1	193	3750	male	2009	0.9577297
327	Chinstrap	Dream	48.1	16.4	199	3325	female	2009	0.6345787

#### 3.1.4.2 Python

```
import pandas as pd
import numpy as np

indexes = np.random.randint(0, len(penguins.index), 5)
```

```
newpenguins = penguins[(penguins.index).isin(indexes)]
newpenguins
```

	randomNumber	rowid	species	...	body_mass_g	sex	year
146	0.765029	147	Adelie	...	4250.0	male	2009
260	0.724817	261	Gentoo	...	4575.0	female	2009
285	0.601037	286	Chinstrap	...	3700.0	male	2007
331	0.747758	332	Chinstrap	...	3450.0	male	2009
340	0.613962	341	Chinstrap	...	3400.0	female	2009

[5 rows x 10 columns]

### 3.1.5 Exercise

Bagi jadi 3 - ambil sampel sebanyak:

1. 40 dari dataset Iris (<https://www.kaggle.com/datasets/arshid/iris-flower-dataset>)
2. 14 dari dataset mtcars (<https://gist.github.com/seankross/a412dfbd88b3db70b74b>)
3. 58 dari dataset penguins (<https://gist.github.com/slopp/ce3b90b9168f2f921784de84fa445651>)

## 3.2 Menduga Parameter populasi

Beberapa rumus yang digunakan untuk menduga parameter populasi:

### 3.2.0.1 Mean

$$\hat{\mu} = \bar{y} = \frac{\sum_{i=1}^n y_i}{n} \quad (\text{mean})$$

$$\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{s^2}{n} \quad (\text{ragam penduga})$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\hat{\mu} \pm 2\sqrt{\hat{V}(\bar{y})} \quad (\text{selang kepercayaan})$$

Penduga bagi mean adalah rata-rata sampel, yaitu total nilai dari sampel dibagi ukuran sampel. Bagaimana untuk ragam penduga? Penurunan rumus ini dimulai dari nilai  $V(\bar{y})$ . Dalam kasus populasi tak hingga yang saling bebas dan memiliki ragam sama: