JOBSHEET 5

PENERAPAN PYTHON DALAM SAMPLING DAN DISTRIBUSI SAMPLING

TUJUAN

1. Mahasiswa mampu menjelaskan perbedaan antara teknik simple random, stratified, systematic, dan cluster sampling.
2. Mahasiswa mampu menerapkan masing-masing teknik sampling menggunakan Python dan pustaka pandas serta numpy.
3. Mahasiswa mampu mengevaluasi hasil sampel dari masing-masing teknik.
4. Memahami konsep distribusi sampling.
5. Melakukan simulasi distribusi sampling menggunakan Python.
6. Menafsirkan hasil distribusi sampling dan membandingkannya dengan distribusi normal.

PENJELASAN UMUM

- Pada contoh menggunakan data array yang di-hard code
- Untuk menjalankan kode python, dapat menggunakan Google Collaboratory atau Visual Studio Code, dengan menginstal python sebelumnya

MATERI 1 - TEKNIK SAMPLING

Bagian 1: Import Library dan Generate Dataset

```python
import pandas as pd
import numpy as np

# Membuat dataset contoh
np.random.seed(42)  # Untuk reproduksibilitas
n = 100
data = {
    'Provinsi': np.random.choice(['Jawa Barat', 'Jawa Tengah', 'Jawa Timur', 'DKI Jakarta'], size=n),
    'Usia': np.random.randint(18, 65, size=n),
    'Pendapatan': np.random.randint(5000000, 50000000, size=n),
    'Pendidikan': np.random.choice(['SMA', 'D1', 'D3', 'S1', 'S2', 'S3'], size=n)
}
df = pd.DataFrame(data)
```

Bagian 2: Simple Random Sampling

Simple Random Sampling adalah metode di mana setiap individu dalam populasi memiliki peluang yang sama untuk terpilih menjadi sampel. Cocok digunakan jika populasi homogen.

```python
# 1. Simple Random Sampling (Pengambilan sampel acak sederhana)
# Mengambil 20 sampel secara acak dari seluruh populasi
n_samples = 20
simple_random_sample = df.sample(n=n_samples, random_state=42) # random_state untuk reproduksibilitas

print("Simple Random Sample:")
print(simple_random_sample)
print("\nPenjelasan: Setiap individu dalam populasi memiliki peluang yang sama untuk terpilih dalam sampel.")
```

Silakan dicoba dan bagaimana hasilnya?

Bagian 3: Systematic Sampling

Systematic Sampling memilih data berdasarkan interval tetap (k). Misalnya, dari daftar yang sudah diacak, ambil setiap ke-5 baris. Syaratnya, data harus terdistribusi merata untuk menghindari bias.

```python
# 2. Systematic Sampling (Pengambilan sampel sistematis)
# Mengambil sampel setiap k individu dalam populasi
k = 5  # Interval sampling
systematic_sample = df.iloc[::k]

print("\nSystematic Sample:")
print(systematic_sample)
print("\nPenjelasan: Mengambil sampel pada interval tertentu. Misalnya, setiap ke-5 individu.")
```

Silakan dicoba, bagaimana hasilnya?

Bagian 4: Stratified Sampling

Stratified Sampling membagi populasi ke dalam kelompok homogen (strata), seperti berdasarkan provinsi. Dari setiap strata, sampel diambil secara acak. Teknik ini menjamin representasi dari setiap strata.

```python
# 3. Stratified Sampling (Pengambilan sampel berlapis)
# Membagi populasi berdasarkan provinsi dan mengambil sampel dari setiap strata
strata = df.groupby('Provinsi')
stratified_sample = strata.apply(lambda x: x.sample(n=int(n_samples/len(strata)), random_state=42))
# jumlah sampel proporsional untuk setiap strata

print("\nStratified Sample:")
print(stratified_sample)
print("\nPenjelasan: Populasi dibagi menjadi beberapa kelompok (strata), kemudian diambil sampel dari setiap strata.")
```

Silakan dicoba, bagaimana hasilnya?

Bagian 5: Cluster Sampling

Cluster Sampling membagi populasi menjadi kelompok-kelompok alami (kluster), seperti wilayah geografis. Kemudian, beberapa kluster dipilih secara acak, dan seluruh anggota dalam kluster tersebut diambil sebagai sampel.

```
# 4. Cluster Sampling (Pengambilan sampel kluster)
# Membagi populasi menjadi kluster,
# kemudian memilih beberapa kluster secara acak dan mengambil semua individu di dalam kluster tersebut.

# Misal kita bagi berdasarkan provinsi sebagai kluster, dan mengambil 2 provinsi
clusters = df['Provinsi'].unique()
selected_clusters = np.random.choice(clusters, size=2, replace=False) # Mengambil 2 provinsi secara random

cluster_sample = df[df['Provinsi'].isin(selected_clusters)]

print("\nCluster Sample:")
print(cluster_sample)
print("\nPenjelasan: Populasi dibagi menjadi kelompok-kelompok (kluster), lalu beberapa kluster dipilih secara acak. ")
```

Silakan dicoba, bagaimana hasilnya?

MATERI 2 - DISTRIBUSI SAMPLING

Bagian 1: Populasi dan Sampel

Langkah pertama adalah membuat data populasi. Misalnya, kita ingin meneliti tinggi badan mahasiswa. Kita anggap tinggi badan terdistribusi normal dengan rata-rata 165 cm dan standar deviasi 10 cm.

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

# 1. Populasi dan Sampel

# Misalnya, kita punya populasi tinggi badan mahasiswa.
populasi_tinggi_badan = np.random.normal(loc=165, scale=10, size=1000)  # 1000 data, rata-rata 165 cm, standar deviasi 10 cm

# Mengambil sampel acak berukuran 30 dari populasi
ukuran_sampel = 30
sampel_tinggi_badan = np.random.choice(populasi_tinggi_badan, size=ukuran_sampel, replace=False)
```

Silakan dicoba, bagaimana hasilnya?

Bagian 2: Simulasi Distribusi Sampling

Selanjutnya, kita melakukan simulasi pengambilan sampel sebanyak 1000 kali, lalu menghitung rata-rata masing-masing sampel.

```python
# 2. Menghitung Distribusi Sampling Rata-Rata Sampel

# Simulasi pengambilan banyak sampel (misalnya, 1000 sampel)
banyak_sampel = 1000
rata_rata_sampel = []

for _ in range(banyak_sampel):
  sampel = np.random.choice(populasi_tinggi_badan, size=ukuran_sampel, replace=False)
  rata_rata_sampel.append(np.mean(sampel))

rata_rata_sampel = np.array(rata_rata_sampel)
```

Silakan dicoba, bagaimana hasilnya?

Bagian 3: Visualisasi Distribusi Sampling

Distribusi dari rata-rata sampel akan divisualisasikan menggunakan histogram, ditambah dengan:

- Garis rata-rata populasi
- Kurva distribusi normal untuk membandingkan

```python
# 3. Visualisasi Distribusi Sampling

plt.figure(figsize=(10, 6))
plt.hist(rata_rata_sampel, bins=30, density=True, alpha=0.7, label='Distribusi Sampling')

# Menambahkan garis vertikal untuk rata-rata populasi
plt.axvline(np.mean(populasi_tinggi_badan), color='red', linestyle='dashed', linewidth=2, label='Rata-rata Populasi')

# Menambahkan kurva distribusi normal untuk membandingkan
rata_rata_distribusi_sampling = np.mean(rata_rata_sampel)
standar_deviasi_distribusi_sampling = np.std(rata_rata_sampel)
x = np.linspace(min(rata_rata_sampel), max(rata_rata_sampel), 100)
plt.plot(x, stats.norm.pdf(x, loc=rata_rata_distribusi_sampling, scale=standar_deviasi_distribusi_sampling),
         'b-', lw=2, label='Distribusi Normal')

plt.xlabel('Rata-rata Tinggi Badan (cm)')
plt.ylabel('Frekuensi')
plt.title('Distribusi Sampling Rata-rata Tinggi Badan')
plt.legend()
plt.show()
```

Silakan dicoba, bagaimana hasilnya?

TUGAS

Berikut ini terdapat dataset:
https://docs.google.com/spreadsheets/d/1TwlnvVV-4AxW8ti1hyca91E3eFmhgRcI/edit?usp=sharing&ouid=105176740327954211268&rtpof=true&sd=true

Berdasarkan dataset tersebut, kerjakanlah beberapa tugas berikut:

1. **Implementasi Teknik Sampling**
   Implementasikan empat teknik sampling berikut pada dataset yang telah disediakan:
   - Simple Random Sampling: ambil 20 data secara acak dari seluruh dataset.
   - Stratified Sampling: gunakan kolom 'Region' sebagai strata, dan ambil 5 data dari masing-masing region.
   - Systematic Sampling: ambil data dengan interval tertentu (misalnya setiap ke-5 baris).
   - Cluster Sampling: pilih 2 region secara acak sebagai klaster, dan ambil semua data dari klaster tersebut.
2. **Analisis Statistik Sampel**
   Untuk setiap teknik sampling yang telah Anda lakukan di atas:
   - Hitung rata-rata usia (Age) dari data sampel.
   - Hitung rata-rata pendapatan (Income) dari data sampel.
3. **Kesimpulan dan Perbandingan**
   Bandingkan hasil dari keempat teknik sampling tersebut:
   - Mana yang menghasilkan sampel paling representatif terhadap populasi
   - Apa kelebihan dan kekurangan masing-masing teknik dari segi keragaman data, efisiensi, dan kemudahan implementasi?
4. **Distribusi Sampling Rata-rata**
   Lakukan simulasi distribusi sampling dengan langkah-langkah berikut:
   1) Ambil **1000 sampel acak**, masing-masing **berukuran 30 data**, dari populasi.
   2) Hitung **rata-rata pendapatan (`Income`)** dari setiap sampel.
   3) Buat **histogram** dari distribusi rata-rata tersebut.
   4) Bandingkan dengan **rata-rata pendapatan populasi**
   5) Pertanyaan:
      - Apa bentuk distribusinya?
      - Apakah distribusi tersebut mendekati **distribusi normal**?
5. **Pengaruh Ukuran Sampel**

   Ulangi langkah-langkah pada soal nomor 1, namun gunakan **ukuran sampel yang berbeda**, yaitu: **10, 50, dan 100**.

   1) Bandingkan bentuk distribusi rata-rata dari ketiga ukuran sampel tersebut.
   2) Amati bagaimana perubahan **standar deviasi** pada distribusi sampling seiring bertambahnya ukuran sampel.
   3) Jelaskan fenomena tersebut berdasarkan **Teorema Limit Tengah (Central Limit Theorem)**.