

STATISTIC AND DATA ANALYSIS

Nama : Handy Chriastian

NIM : 241012000043

Kelas : 02MKMM001

Soal no 1.

Dengan menggunakan dataset

<https://www.kaggle.com/datasets/akashnath29/lung-cancer-dataset>

Yaitu dataset tentang kanker paru-paru

```
import pandas as pd
import numpy as np
from scipy import stats

# Load data
df = pd.read_csv('dataset_cancer_paru.csv')

# Pilih kolom numerik
kolom_numerik = df.select_dtypes(include=[np.number]).columns.tolist()

# Fungsi analisis pemusatan data
def analisis_pemusatan(df, kolom):
    hasil = []
    for kol in kolom:
        data = df[kol].dropna()
        mean = np.mean(data)
        median = np.median(data)
        mode = stats.mode(data, keepdims=True)[0][0]
        hasil.append({
            'Variabel': kol,
            'Mean': round(mean, 2),
            'Median': round(median, 2),
            'Modus': mode
        })
    return pd.DataFrame(hasil)

# Jalankan analisis dan tampilkan hasil
hasil_analisis = analisis_pemusatan(df, kolom_numerik)
print(hasil_analisis)
```

STATISTIC AND DATA ANALYSIS

Nama : Handy Chriastian

NIM : 241012000043

Kelas : 02MKMM001

```
# Fungsi analisis penyebaran data
def analisis_penyebaran(df, kolom):
    hasil = []
    for kol in kolom:
        data = df[kol].dropna()
        std_dev = np.std(data, ddof=1)
        variance = np.var(data, ddof=1)
        range_data = np.ptp(data)
        hasil.append({
            'Variabel': kol,
            'Standar Deviasi': round(std_dev, 2),
            'Varians': round(variance, 2),
            'Range': round(range_data, 2)
        })
    return pd.DataFrame(hasil)

# Jalankan analisis penyebaran dan tampilkan hasil
hasil_penyebaran = analisis_penyebaran(df, kolom_numerik)
print(hasil_penyebaran)

# Fungsi analisis pendugaan parameter rata-rata
def analisis_pendugaan_rata_rata(df, kolom, alpha=0.01):
    hasil = []
    z_score = stats.norm.ppf(1 - alpha / 2) # Z-score untuk tingkat signifikan
    for kol in kolom:
        data = df[kol].dropna()
        mean = np.mean(data)
        std_dev = np.std(data, ddof=1)
        n = len(data)
        margin_of_error = z_score * (std_dev / np.sqrt(n))
        lower_bound = mean - margin_of_error
        upper_bound = mean + margin_of_error
        hasil.append({
            'Variabel': kol,
            'Rata-rata': round(mean, 2),
            'Batas Bawah': round(lower_bound, 2),
            'Batas Atas': round(upper_bound, 2)
        })
    return pd.DataFrame(hasil)
```

STATISTIC AND DATA ANALYSIS

Nama : Handy Christian

NIM : 241012000043

Kelas : 02MKMM001

```
# Jalankan analisis pendugaan rata-rata dan tampilkan hasil
hasil_pendugaan = analisis_pendugaan_rata_rata(df, kolom_numerik)
print(hasil_pendugaan)
```

a. Pemusatan data

semester2 % /usr/local/bin/python3 /Users/macbook/Documents/semester2/uts_statistik.py				
	Variabel	Mean	Median	Modus
0	AGE	55.17	55.0	54
1	SMOKING	1.49	1.0	1
2	YELLOW_FINGERS	1.51	2.0	2
3	ANXIETY	1.49	1.0	1
4	PEER_PRESSURE	1.50	1.0	1
5	CHRONIC_DISEASE	1.51	2.0	2
6	FATIGUE	1.49	1.0	1
7	ALLERGY	1.51	2.0	2
8	WHEEZING	1.50	1.0	1
9	ALCOHOL_CONSUMING	1.49	1.0	1
10	COUGHING	1.51	2.0	2
11	SHORTNESS_OF_BREATH	1.49	1.0	1
12	SWALLOWING_DIFFICULTY	1.49	1.0	1
13	CHEST_PAIN	1.50	1.0	1
	Variabel	Standar Deviasi	Varians	Range
0	AGE	14.72	216.79	50
1	SMOKING	0.50	0.25	1
2	YELLOW_FINGERS	0.50	0.25	1
3	ANXIETY	0.50	0.25	1
4	PEER_PRESSURE	0.50	0.25	1
5	CHRONIC_DISEASE	0.50	0.25	1
6	FATIGUE	0.50	0.25	1
7	ALLERGY	0.50	0.25	1
8	WHEEZING	0.50	0.25	1
9	ALCOHOL_CONSUMING	0.50	0.25	1
10	COUGHING	0.50	0.25	1
11	SHORTNESS_OF_BREATH	0.50	0.25	1
12	SWALLOWING_DIFFICULTY	0.50	0.25	1
13	CHEST_PAIN	0.50	0.25	1

- **AGE** (usia): nilai rata-rata 55.17, median 55, dan modus 54. Variabel lain seperti **SMOKING**, **ANXIETY**, dan **CHEST_PAIN** memiliki nilai antara 1 dan 2, menunjukkan data biner (misalnya: 1 = tidak, 2 = ya). Sebagian besar variabel memiliki median dan modus = 1 atau 2, menandakan kecenderungan jawaban dominan pada kategori tertentu.

STATISTIC AND DATA ANALYSIS

Nama : Handy Chriastian

NIM : 241012000043

Kelas : 02MKMM001

b. Penyebaran data

	Variabel	Standar Deviasi	Varians	Range
0	AGE	14.72	216.79	50
1	SMOKING	0.50	0.25	1
2	YELLOW_FINGERS	0.50	0.25	1
3	ANXIETY	0.50	0.25	1
4	PEER_PRESSURE	0.50	0.25	1
5	CHRONIC_DISEASE	0.50	0.25	1
6	FATIGUE	0.50	0.25	1
7	ALLERGY	0.50	0.25	1
8	WHEEZING	0.50	0.25	1
9	ALCOHOL_CONSUMING	0.50	0.25	1
10	COUGHING	0.50	0.25	1
11	SHORTNESS_OF_BREATH	0.50	0.25	1
12	SWALLOWING_DIFFICULTY	0.50	0.25	1
13	CHEST_PAIN	0.50	0.25	1

- **AGE** memiliki penyebaran besar (standar deviasi 14.72, range 50), menunjukkan variasi umur yang luas.

Variabel lain (seperti **SMOKING**, **ANXIETY**, dll.) memiliki penyebaran kecil (standar deviasi 0.50, range 1) karena berupa data biner (1 dan 2).

Artinya, hanya **AGE** yang kontinu; sisanya bersifat kategorikal.

c. analisis pendugaan parameter rata-rata, dengan tingkat signifikan 99%.

	Variabel	Rata-rata	Batas Bawah	Batas Atas
0	AGE	55.17	54.48	55.86
1	SMOKING	1.49	1.47	1.51
2	YELLOW_FINGERS	1.51	1.49	1.54
3	ANXIETY	1.49	1.47	1.52
4	PEER_PRESSURE	1.50	1.48	1.52
5	CHRONIC_DISEASE	1.51	1.49	1.53
6	FATIGUE	1.49	1.47	1.51
7	ALLERGY	1.51	1.48	1.53
8	WHEEZING	1.50	1.47	1.52
9	ALCOHOL_CONSUMING	1.49	1.47	1.51
10	COUGHING	1.51	1.49	1.53
11	SHORTNESS_OF_BREATH	1.49	1.46	1.51
12	SWALLOWING_DIFFICULTY	1.49	1.47	1.51
13	CHEST_PAIN	1.50	1.48	1.52

AGE memiliki rata-rata 55.17 dengan interval kepercayaan 99% antara **54.48 – 55.86** → artinya, kita 99% yakin bahwa rata-rata umur populasi berada dalam rentang ini.

Variabel lain seperti **SMOKING**, **ANXIETY**, dll. memiliki interval sempit (misal: 1.47 – 1.51), karena data biner.

STATISTIC AND DATA ANALYSIS

Nama : Handy Chriastian

NIM : 241012000043

Kelas : 02MKMM001

Rentang sempit menandakan data stabil dan varians kecil. **Kesimpulan:** AGE menunjukkan variasi yang signifikan, sementara variabel biner cenderung konsisten.

Soal no. 2

Program Python

```
import math
from scipy.stats import norm

# Diketahui
n = 200          # ukuran sampel
x_bar = 1.8      # rata-rata sampel
mu_0 = 2.0       # rata-rata hipotesis nol
sigma = 0.5      # standar deviasi
alpha = 0.05     # tingkat signifikansi

# (a) Hipotesis
print("Hipotesis:")
print("H0:  $\mu \geq 2$  (waktu respon tidak kurang dari 2 detik)")
print("H1:  $\mu < 2$  (waktu respon kurang dari 2 detik)")

# (b) Hitung nilai statistik uji
z = (x_bar - mu_0) / (sigma / math.sqrt(n))
print(f"\n(b) Nilai statistik uji (z): {z:.2f}")

# (c) Nilai kritis dan p-value
z_kritis = norm.ppf(alpha) # untuk uji satu sisi kiri
p_value = norm.cdf(z)

print(f"(c) Nilai z kritis: {z_kritis:.2f}")
print(f"    p-value      : {p_value:.8f}")

# (d) Kesimpulan
print("\n(d) Kesimpulan:")
if z < z_kritis:
    print("Tolak H0: Ada cukup bukti bahwa rata-rata waktu respon < 2 detik.")
else:
    print("Gagal tolak H0: Tidak ada cukup bukti bahwa rata-rata waktu respon < 2 detik.")
```

STATISTIC AND DATA ANALYSIS

Nama : Handy Chriastian

NIM : 241012000043

Kelas : 02MKMM001

Output :

```
Welcome  uts_statistik.py M  menguji_hipotesis.py U X
menguji_hipotesis.py > ...
1  import math
2  from scipy.stats import norm
3
4  # Diketahui
5  n = 200          # ukuran sampel
6  x_bar = 1.8      # rata-rata sampel
7  mu_0 = 2.0       # rata-rata hipotesis nol
8  sigma = 0.5      # standar deviasi
9  alpha = 0.05     # tingkat signifikansi
10
11 # (a) Hipotesis
12 print("Hipotesis:")
13 print("H0:  $\mu \geq 2$  (waktu respon tidak kurang dari 2 detik)")
14 print("H1:  $\mu < 2$  (waktu respon kurang dari 2 detik)")
15
16 # (b) Hitung nilai statistik uji
17 z = (x_bar - mu_0) / (sigma / math.sqrt(n))
18 print(f"\n(b) Nilai statistik uji (z): {z:.2f}")
19
20 # (c) Nilai kritis dan p-value
21 z_kritis = norm.ppf(alpha) # untuk uji satu sisi kiri
22
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS  ROBOT DOCUMENTATION  ROBOT OUTPUT
/usr/local/bin/python3 /Users/macbook/Documents/semester2/menguji_hipotesis.py
macbook@Handys-Macbook-Pro semester2 % /usr/local/bin/python3 /Users/macbook/Documents/semester2/menguji_hipotesis.py
Hipotesis:
H0:  $\mu \geq 2$  (waktu respon tidak kurang dari 2 detik)
H1:  $\mu < 2$  (waktu respon kurang dari 2 detik)

(b) Nilai statistik uji (z): -5.66
(c) Nilai z kritis: -1.64
    p-value       : 0.00000001

(d) Kesimpulan:
Tolak H0: Ada cukup bukti bahwa rata-rata waktu respon < 2 detik.
macbook@Handys-Macbook-Pro semester2 %
```

Hipotesis:

H0: $\mu \geq 2$ (waktu respon tidak kurang dari 2 detik)

H1: $\mu < 2$ (waktu respon kurang dari 2 detik)

(b) Nilai statistik uji (z): -5.66

(c) Nilai z kritis: -1.64

p-value : 0.00000001

(d) Kesimpulan:

Tolak H0: Ada cukup bukti bahwa rata-rata waktu respon < 2 detik.

STATISTIC AND DATA ANALYSIS

Nama : Handy Chriastian

NIM : 241012000043

Kelas : 02MKMM001

Soal no. 3

```
probabilitas.py > ...
1  from scipy.stats import norm
2
3  # Diketahui
4  mu = 200      # mean (rata-rata) dalam milidetik
5  sigma = 50    # standar deviasi dalam milidetik
6
7  # (a) Probabilitas waktu respons kurang dari 250 milidetik
8  x_a = 250
9  prob_a = norm.cdf(x_a, loc=mu, scale=sigma)
10
11 # (b) Probabilitas waktu respons antara 200 dan 270 milidetik
12 x_b1 = 200
13 x_b2 = 270
14 prob_b = norm.cdf(x_b2, loc=mu, scale=sigma) - norm.cdf(x_b1, loc=mu, scale=sigma)
15
16 print(f"(a) Probabilitas waktu respons kurang dari {x_a} ms: {prob_a:.4f} ({prob_a*100:.2f}%)")
17 print(f"(b) Probabilitas waktu respons antara {x_b1} ms dan {x_b2} ms: {prob_b:.4f} ({prob_b*100:.2f}%)")
18 # (c) Probabilitas waktu respons lebih dari 300 milidetik

PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS  ROBOT DOCUMENTATION  ROBOT OUTPUT

/usr/local/bin/python3 /Users/macbook/Documents/semester2/probabilitas.py
macbook@Handys-Macbook-Pro semester2 % /usr/local/bin/python3 /Users/macbook/Documents/semester2/probabilitas.py
(a) Probabilitas waktu respons kurang dari 250 ms: 0.8413 (84.13%)
(b) Probabilitas waktu respons antara 200 ms dan 270 ms: 0.4192 (41.92%)
macbook@Handys-Macbook-Pro semester2 %
```

(a) Probabilitas waktu respons kurang dari 250 milidetik

Kita ingin mencari probabilitas bahwa waktu respons $X < 250$ milidetik.

Langkah pertama adalah menghitung nilai Z untuk $X = 250$:

$$Z = \frac{250 - 200}{50} = 1$$

Kemudian, kita dapat mencari probabilitas kumulatif untuk nilai $Z = 1$ dari tabel distribusi normal standar atau menggunakan fungsi distribusi kumulatif normal (CDF). Nilai CDF untuk $Z = 1$ adalah sekitar 0.8413.

Jadi, probabilitas waktu respons kurang dari 250 milidetik adalah sekitar 0.8413 atau 84.13%.

(b) Probabilitas waktu respons antara 200 dan 270 milidetik

STATISTIC AND DATA ANALYSIS

Nama : Handy Chriastian

NIM : 241012000043

Kelas : 02MKMM001

Untuk mencari probabilitas waktu respons berada di antara 200 milidetik dan 270 milidetik, kita akan menghitung nilai Z untuk kedua waktu respons tersebut.

1. Untuk $X=200$ $X = 200$ milidetik, kita sudah tahu bahwa:
 $Z = \frac{200 - 200}{50} = 0$
Probabilitas kumulatif untuk $Z=0$ adalah 0.5.
2. Untuk $X=270$ $X = 270$ milidetik, kita menghitung nilai Z:
 $Z = \frac{270 - 200}{50} = 1.4$
Probabilitas kumulatif untuk $Z=1.4$ adalah sekitar 0.9192.

untuk mendapatkan probabilitas antara 200 milidetik dan 270 milidetik, kita menghitung selisih antara probabilitas kumulatif $Z=1.4$ dan $Z=0$:

$$P(200 \leq X \leq 270) = P(Z=1.4) - P(Z=0) = 0.9192 - 0.5 = 0.4192$$
$$P(200 \leq X \leq 270) = P(Z=1.4) - P(Z=0) = 0.9192 - 0.5 = 0.4192$$

Jadi, probabilitas bahwa waktu respons server berada di antara 200 milidetik dan 270 milidetik adalah sekitar 0.4192 atau 41.92%.