Borcelle

Data & Sampling Dustribution

Dipresentasikan Oleh: Asthagina Delia Putri (20244920001)

www.reallygreatsite.com

27 June, 2024

Data Sains

DISTRIBUSI NORMAL

Distribusi normal adalah distribusi probabilitas kontinu yang berbentuk lonceng dan simetris terhadap rata-rata. Digunakan dalam berbagai analisis statistik.

pdf :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \qquad -\infty < x < \infty$$

cdf :

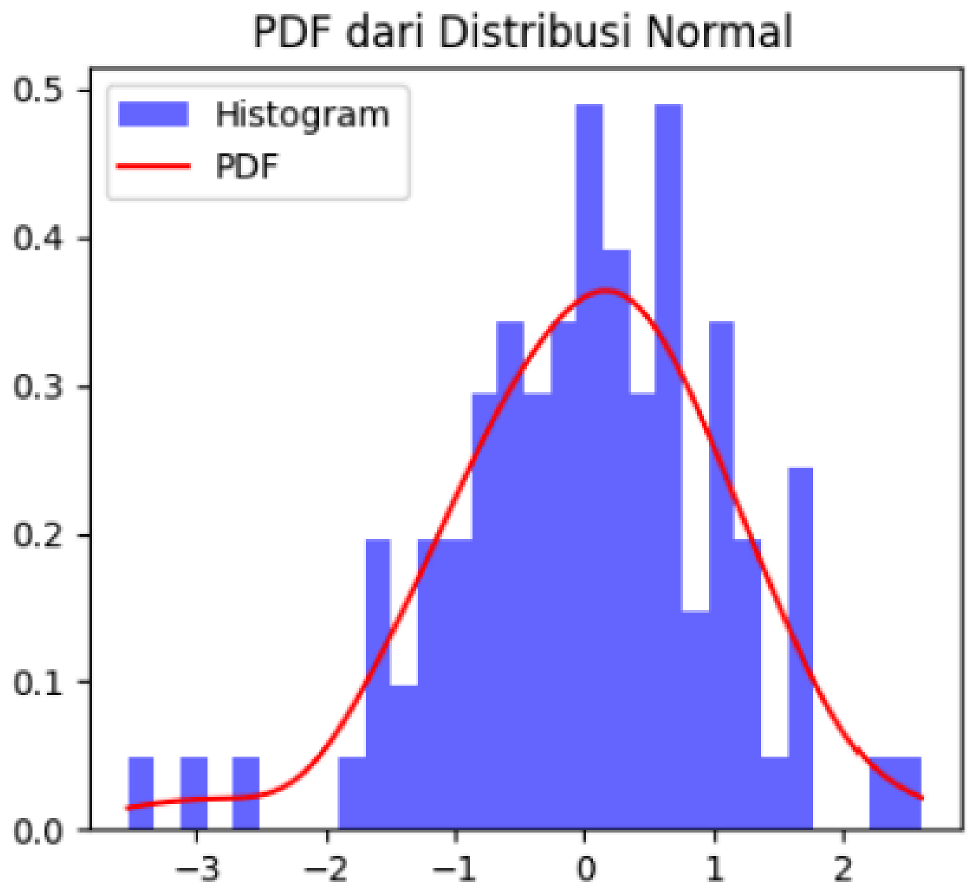
$$F(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

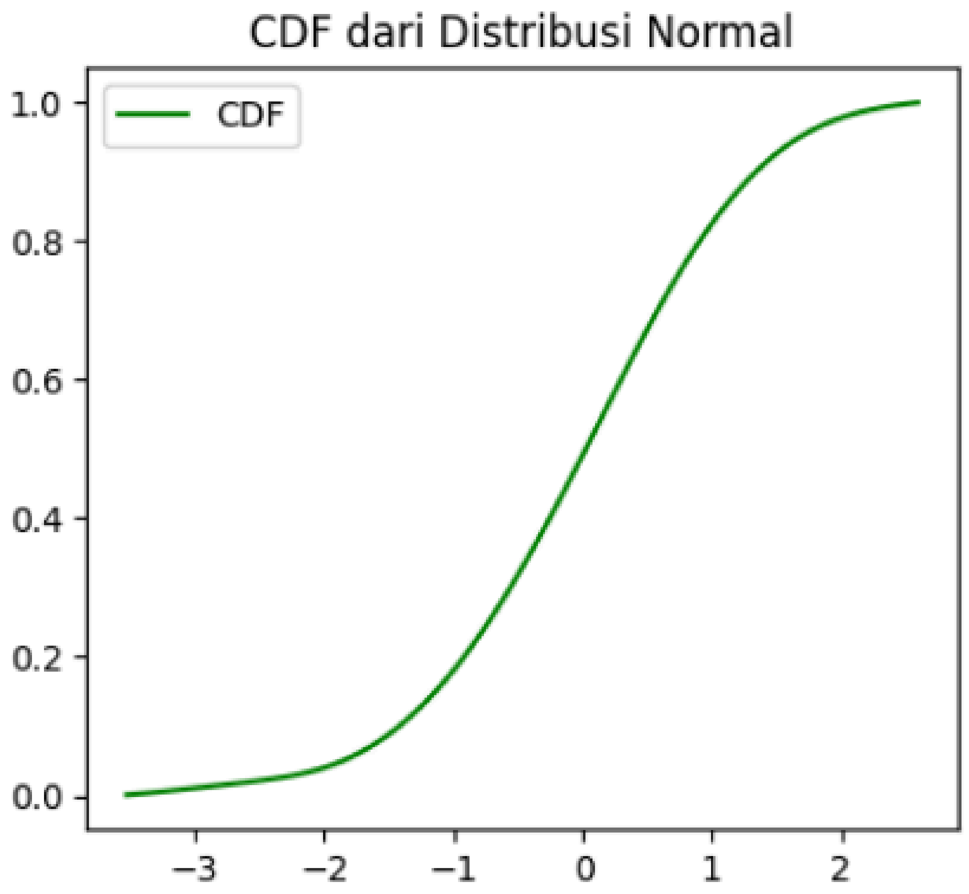
kode phyton data :

```
distribution = {
    "Normal": stats.norm.rvs(loc=0, scale=1, size=100),
}

for name, data in distribution.items():
    x = np.linspace(min(data), max(data), 100)
    pdf = stats.gaussian_kde(data).evaluate(x)  # Estimasi PDF
    cdf = np.cumsum(pdf) / sum(pdf)  # Estimasi CDF

    mean = np.mean(data)
    std = np.std(data)

    print(f"Distribusi {name}: Mean = {mean :.4f}, Std = {std:.4f}")

    plt.figure(figsize=(10, 4))
    plt.subplot(1, 2, 1)
    plt.hist(data, bins=30, density=True, alpha=0.6, color='b',
label='Histogram')
    plt.plot(x, pdf, color='red', label='PDF')
    plt.title(f"PDF dari Distribusi {name}")
    plt.legend()

    plt.subplot(1, 2, 2)
    plt.plot(x, cdf, color='green', label='CDF')
    plt.title(f"CDF dari Distribusi {name}")
    plt.legend()
    plt.show()
```

Data Sains

Visualisasi data :





Distribusi Normal : Mean = 0.0049, Std = 1.0660

Sains Data

DISTRIBUSI LONG TAILED

Distribusi dengan ekor panjang menunjukkan probabilitas tinggi untuk nilai ekstrem dibandingkan distribusi norma

pdf :

$$f(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, \quad x \geq x_m$$

cdf :

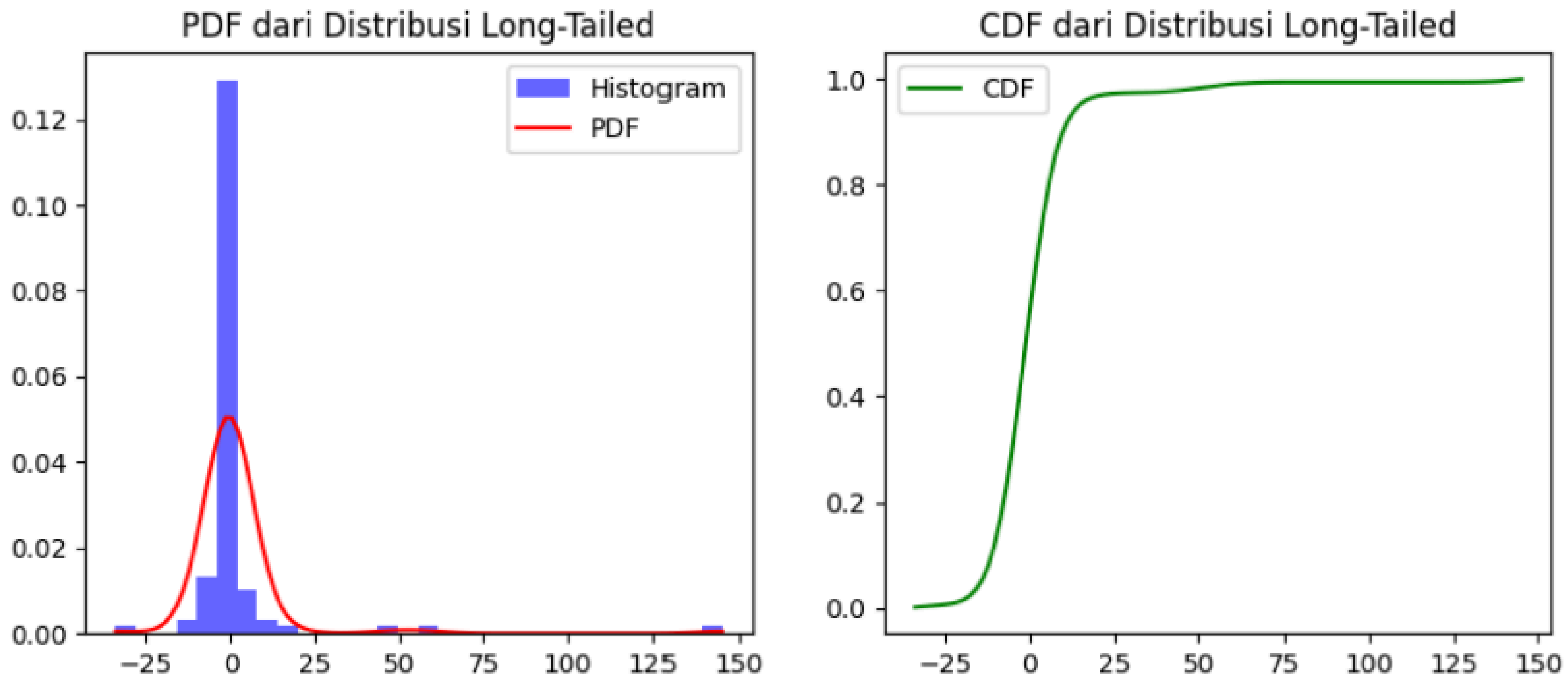
$$F(x) = 1 - \left(\frac{x_m}{x}\right)^\alpha, \quad x \geq x_m$$

kode phyton data :

```
distribution = {
    "Long-Tailed": stats.cauchy.rvs(size=100)
}

for name, data in distribution.items():
    x = np.linspace(min(data), max(data), 100)
    pdf = stats.gaussian_kde(data).evaluate(x)  # Estimasi PDF
    cdf = np.cumsum(pdf) / sum(pdf)  # Estimasi CDF
    mean = np.mean(data)
    std = np.std(data)

    print(f"Distribusi {name}: Mean = {mean :.4f}, Std = {std:.4f}")

    plt.figure(figsize=(10, 4))
    plt.subplot(1, 2, 1)
    plt.hist(data, bins=30, density=True, alpha=0.6, color='b',
label='Histogram')
    plt.plot(x, pdf, color='red', label='PDF')
    plt.title(f"PDF dari Distribusi {name}")
    plt.legend()

    plt.subplot(1, 2, 2)
    plt.plot(x, cdf, color='green', label='CDF')
    plt.title(f"CDF dari Distribusi {name}")
    plt.legend()

    plt.show()
```

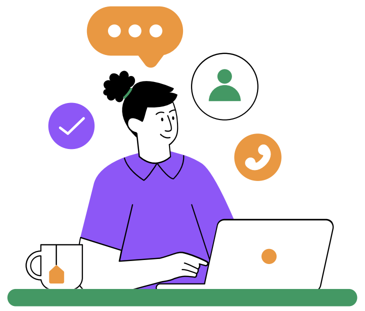
Sains Data

visualisasi data :



Distribusi Long-Tailed : Mean = 1.7525, Std = 17.0255

Sains Data



DISTRIBUSI STUDENT'S T

Digunakan dalam pengujian hipotesis untuk sampel kecil.

kode phyton data :

```
distribution = {
    "Student's t": stats.t.rvs(df=10, size=100)
}

for name, data in distribution.items():
    x = np.linspace(min(data), max(data), 100)
    pdf = stats.gaussian_kde(data).evaluate(x)  # Estimasi PDF
    cdf = np.cumsum(pdf) / sum(pdf)  # Estimasi CDF

    mean = np.mean(data)
    std = np.std(data)

    print(f"Distribusi {name}: Mean = {mean :.4f}, Std = {std:.4f}")

    plt.figure(figsize=(10, 4))
    plt.subplot(1, 2, 1)
    plt.hist(data, bins=30, density=True, alpha=0.6, color='b',
label='Histogram')
    plt.plot(x, pdf, color='red', label='PDF')
    plt.title(f"PDF dari Distribusi {name}")
    plt.legend()

    plt.subplot(1, 2, 2)
    plt.plot(x, cdf, color='green', label='CDF')
    plt.title(f"CDF dari Distribusi {name}")
    plt.legend()

    plt.show()
```

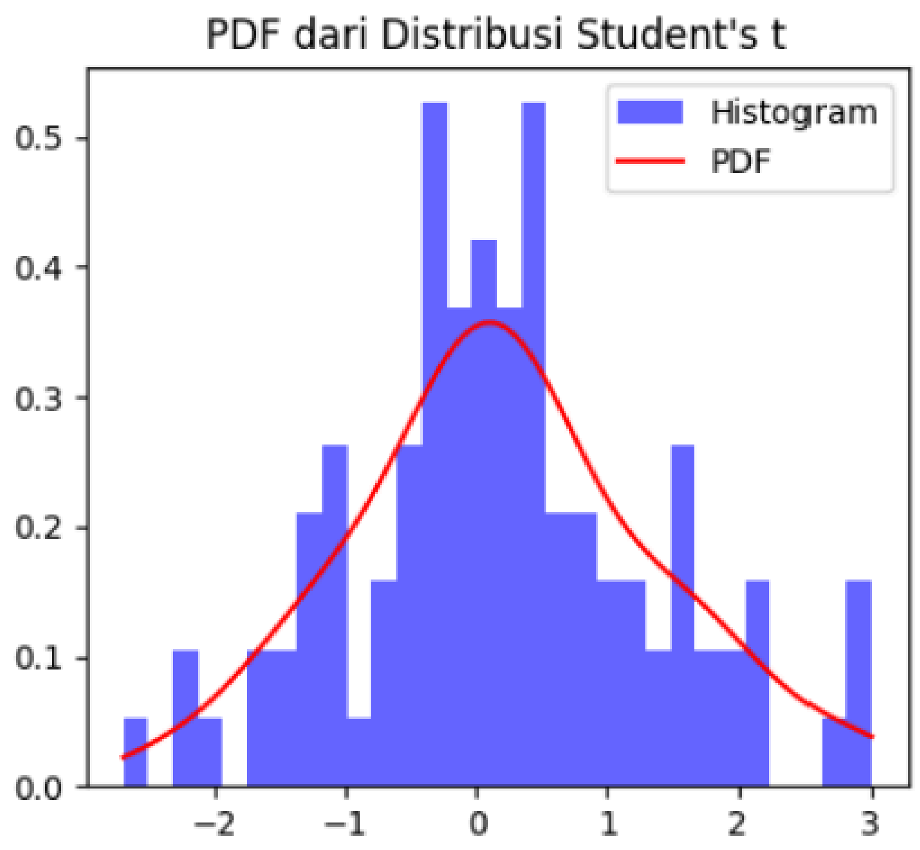
pdf :

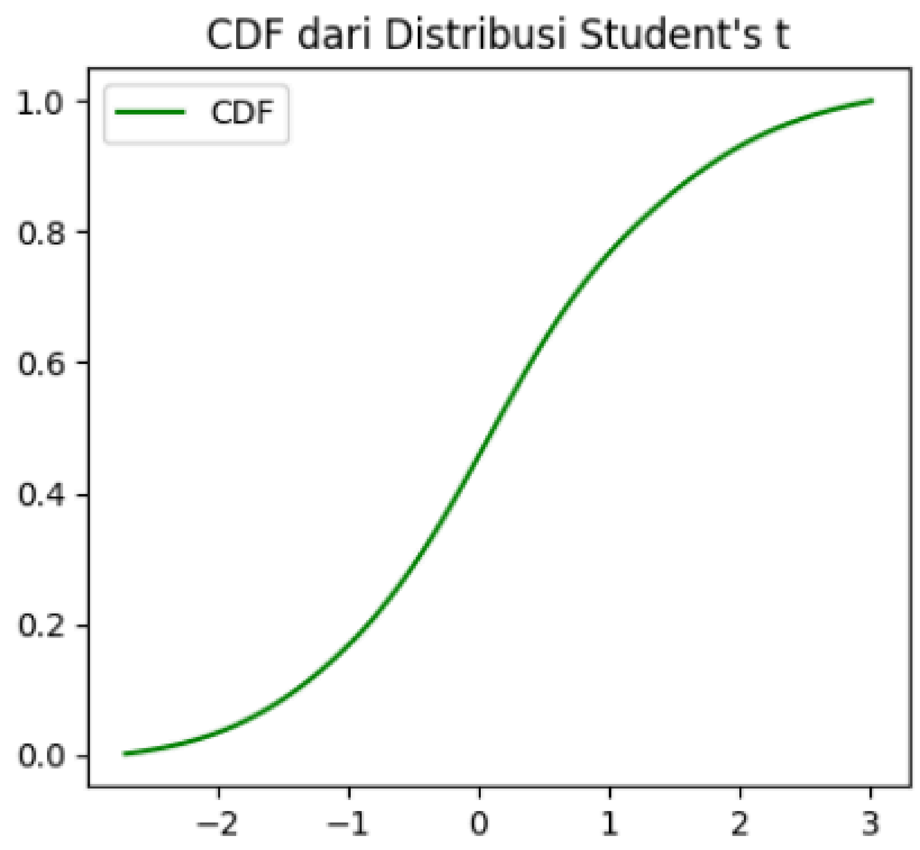
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

cdf :

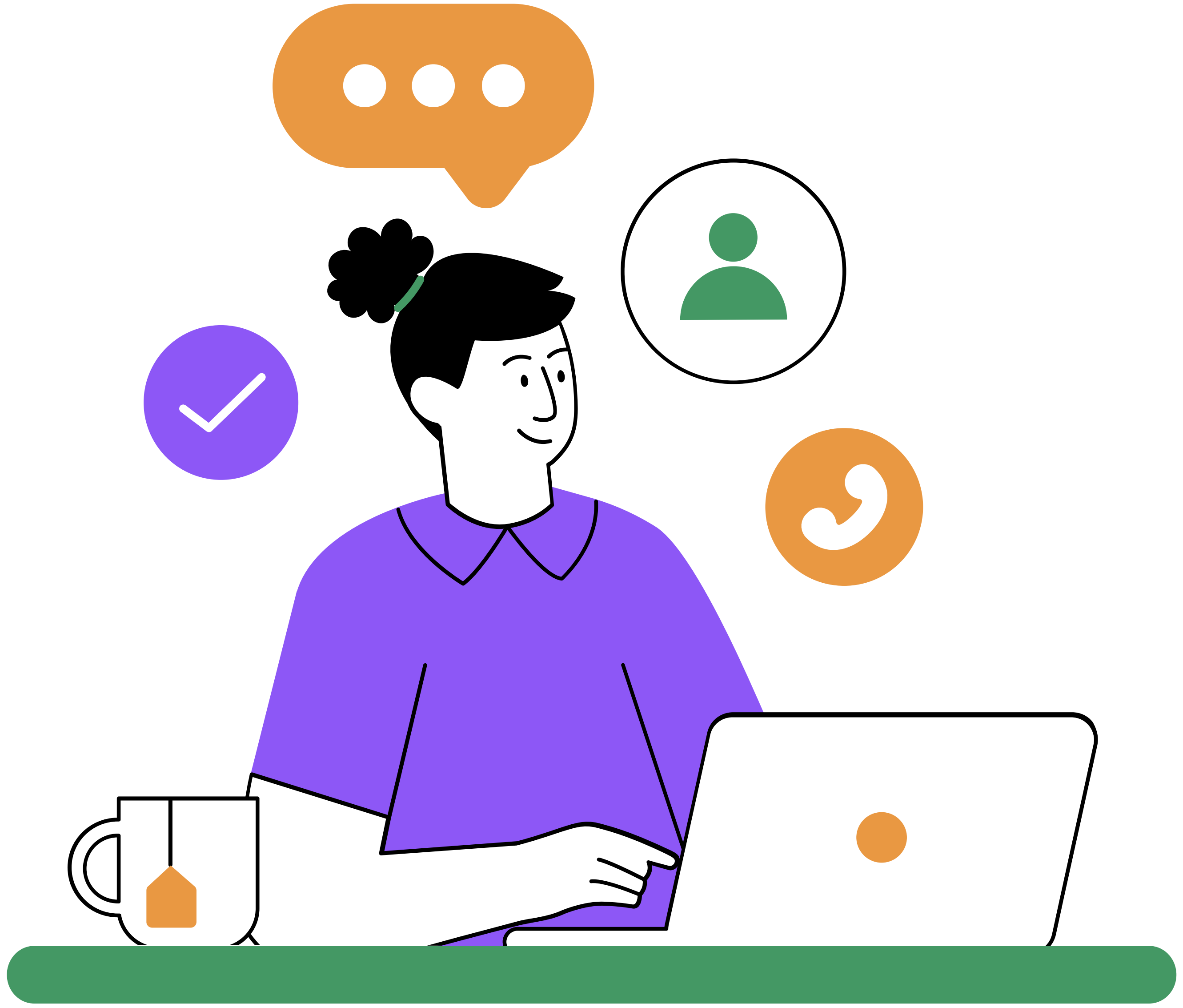
$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

visualisasi data :





Distribusi Student's t : Mean = 0.2042, Std = 1.1656



Sains Data

DISTRIBUSI BINOMIAL

Digunakan untuk menghitung probabilitas sukses dalam percobaan Bernoulli yang berulang.

kode phyton data :

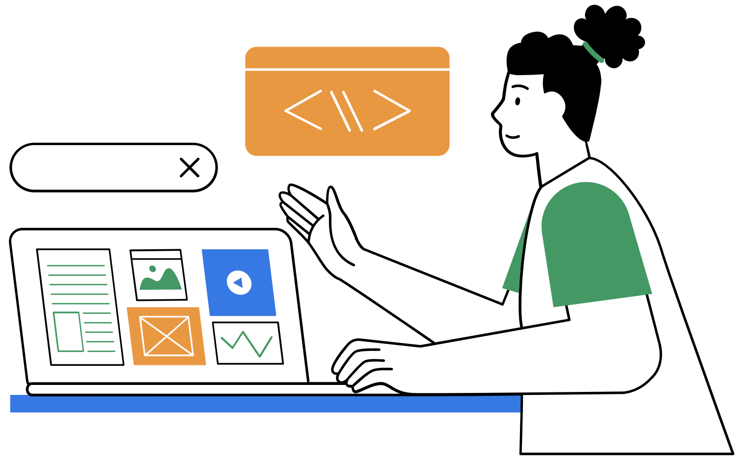
```
distribution = {
    "binomial": np.random.binomial(n=10, p=0.5, size=100)
}

for name, data in distribution.items():
    x = np.linspace(min(data), max(data), 100)
    pdf = stats.gaussian_kde(data).evaluate(x)  # Estimasi PDF
    cdf = np.cumsum(pdf) / sum(pdf)  # Estimasi CDF

    mean = np.mean(data)
    std = np.std(data)

    print(f"Distribusi {name}: Mean = {mean :.4f}, Std = {std:.4f}")

    plt.figure(figsize=(10, 4))
    plt.subplot(1, 2, 1)
    plt.hist(data, bins=30, density=True, alpha=0.6, color='b',
label='Histogram')
    plt.plot(x, pdf, color='red', label='PDF')
    plt.title(f"PDF dari Distribusi {name}")
    plt.legend()

    plt.subplot(1, 2, 2)
    plt.plot(x, cdf, color='green', label='CDF')
    plt.title(f"CDF dari Distribusi {name}")
    plt.legend()

    plt.show()
```

pdf :

$$P(x = k) = C^n_k p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

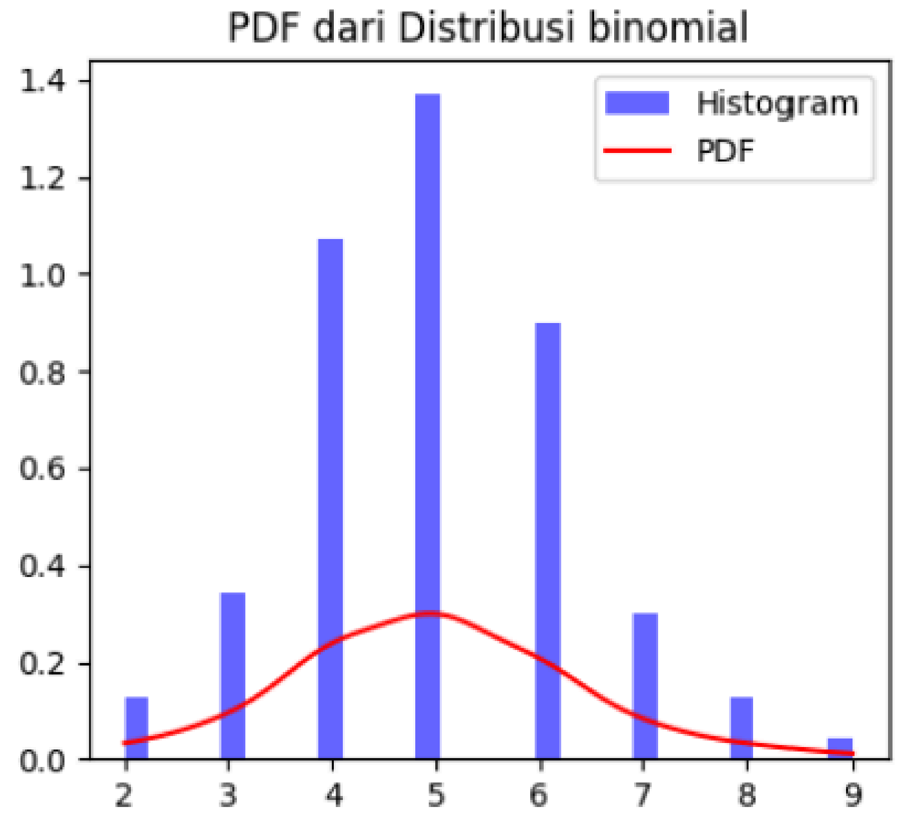
cdf :

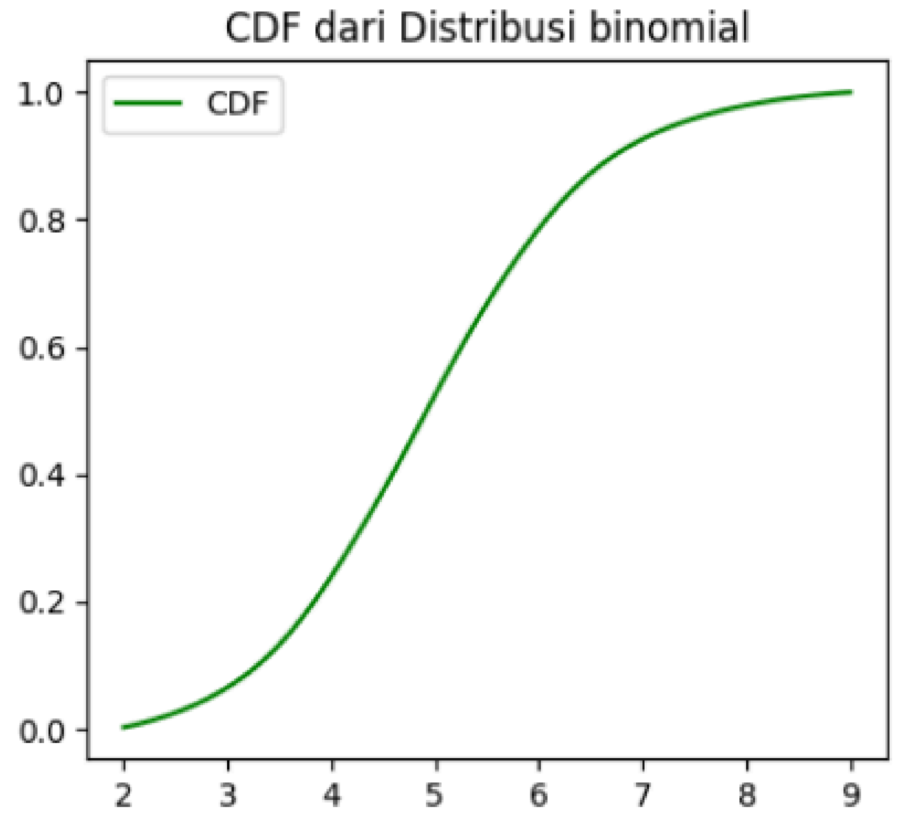
$$F(x) = \sum_{k=0}^{x} P(X = k)$$



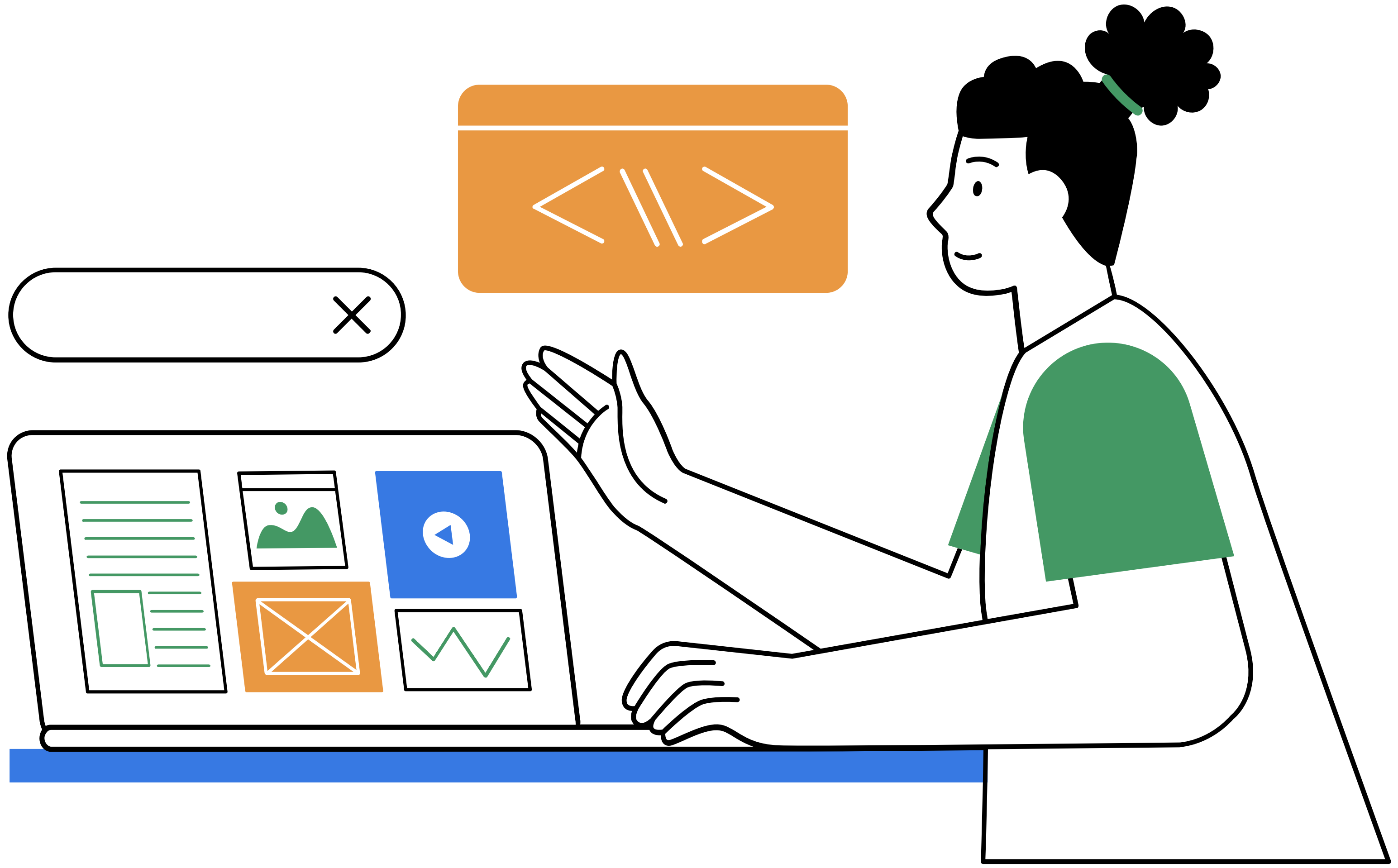
Sains Data

visualisasi data :





Distribusi binomial : Mean = 4.9800, Std = 1.3265



Sains Data

DISTRIBUSI CHI-SQUARE

Digunakan dalam uji statistik untuk varians populasi dan goodness-of-fit test.

pdf :

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

cdf :

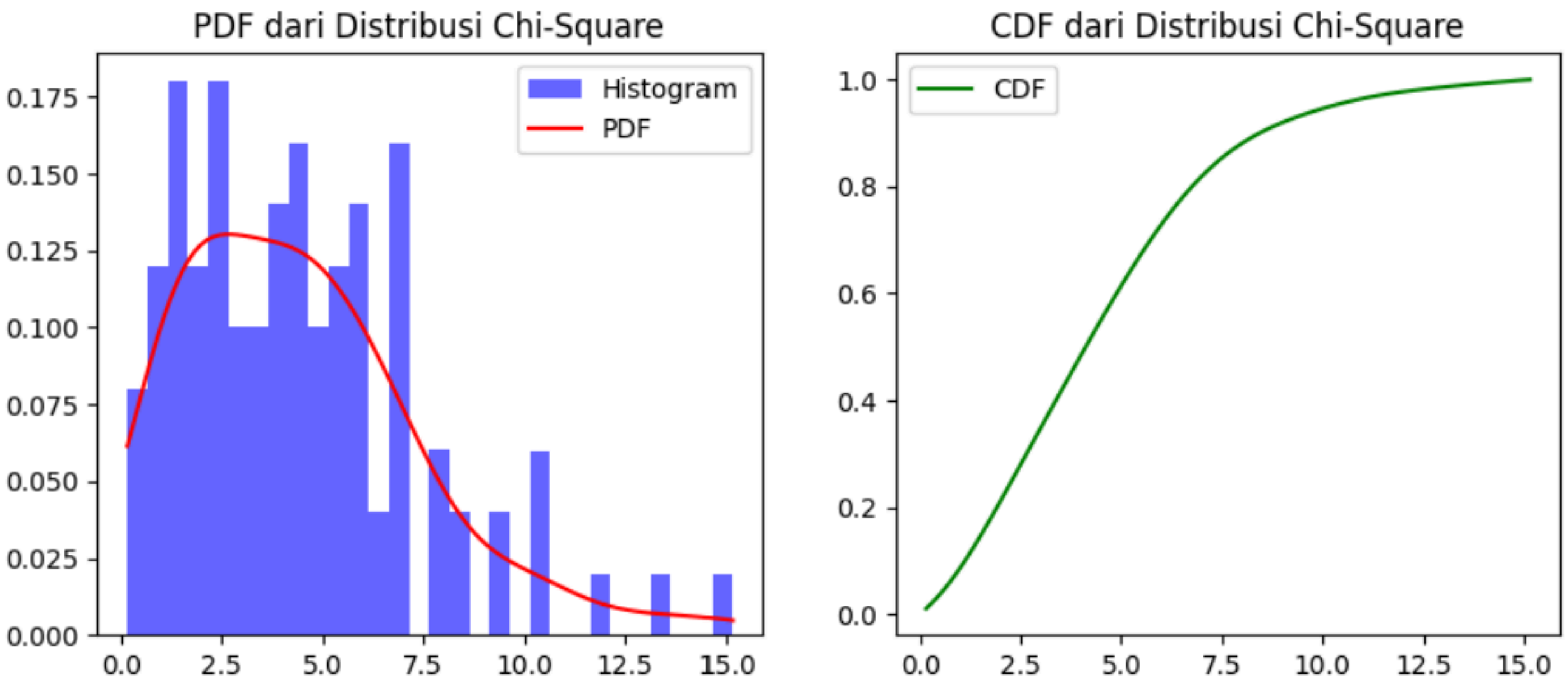
$$F(x) = \frac{1}{\Gamma(k/2)}\gamma\left(\frac{k}{2}, \frac{x}{2}\right)$$

kode phyton data :

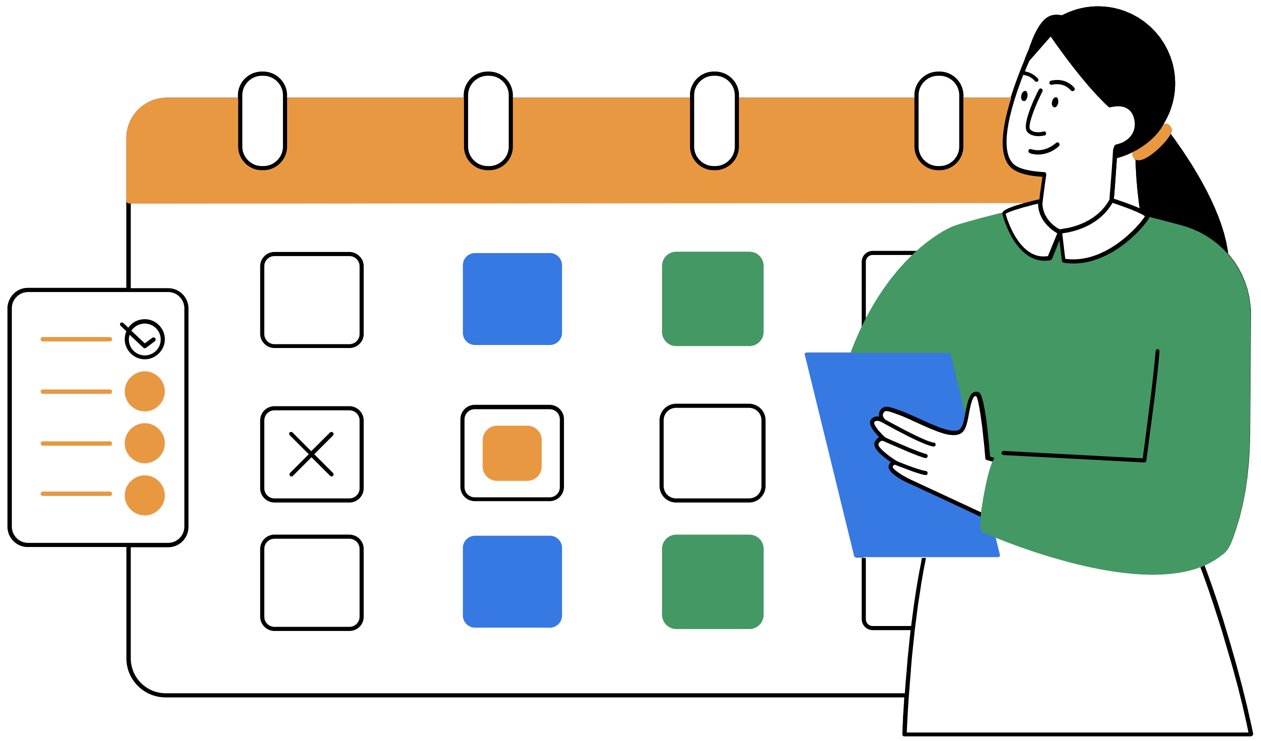
```python
distribution = {
    "Chi-Square": stats.chi2.rvs(df=4, size=100)
}

for name, data in distribution.items():
    x = np.linspace(min(data), max(data), 100)
    pdf = stats.gaussian_kde(data).evaluate(x)  # Estimasi PDF
    cdf = np.cumsum(pdf) / sum(pdf)  # Estimasi CDF

    mean = np.mean(data)
    std = np.std(data)

    print(f"Distribusi {name}: Mean = {mean :.4f}, Std = {std:.4f}")

    plt.figure(figsize=(10, 4))
    plt.subplot(1, 2, 1)
    plt.hist(data, bins=30, density=True, alpha=0.6, color='b',
label='Histogram')
    plt.plot(x, pdf, color='red', label='PDF')
    plt.title(f"PDF dari Distribusi {name}")
    plt.legend()

    plt.subplot(1, 2, 2)
    plt.plot(x, cdf, color='green', label='CDF')
    plt.title(f"CDF dari Distribusi {name}")
    plt.legend()

    plt.show()
```

Sains Data

visualisasi data :

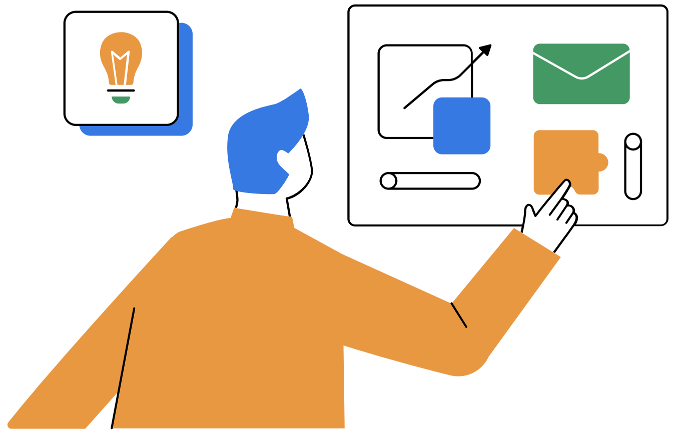


Distribusi Chi-Square : Mean = 4.4182, Std = 2.9563



Sains Data

DISTRIBUSI F



Distribusi untuk membandingkan dua varians dalam analisis varians (ANOVA).

pdf :

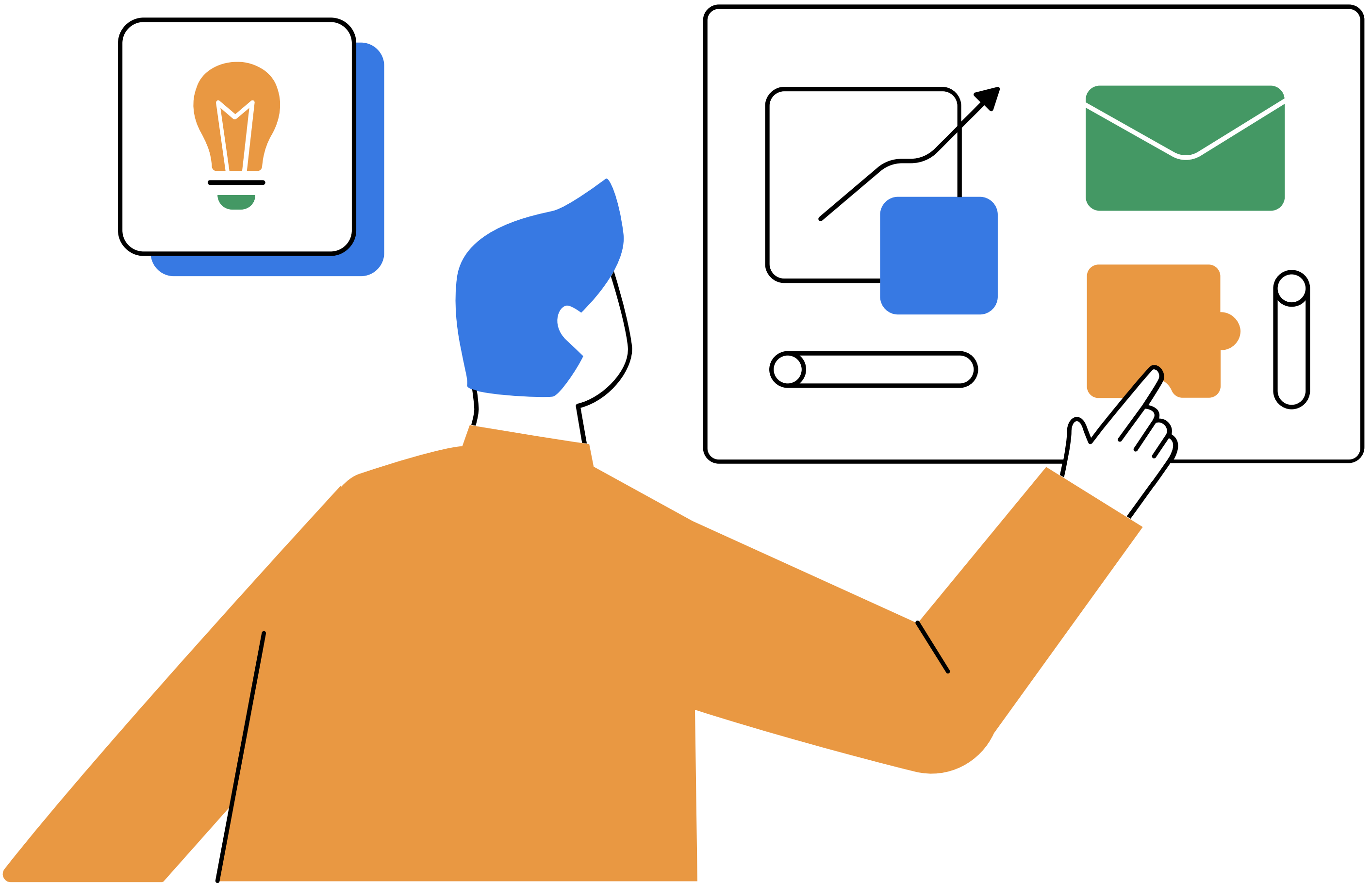
$$F = \frac{s_1^2}{s_2^2}$$

cdf :

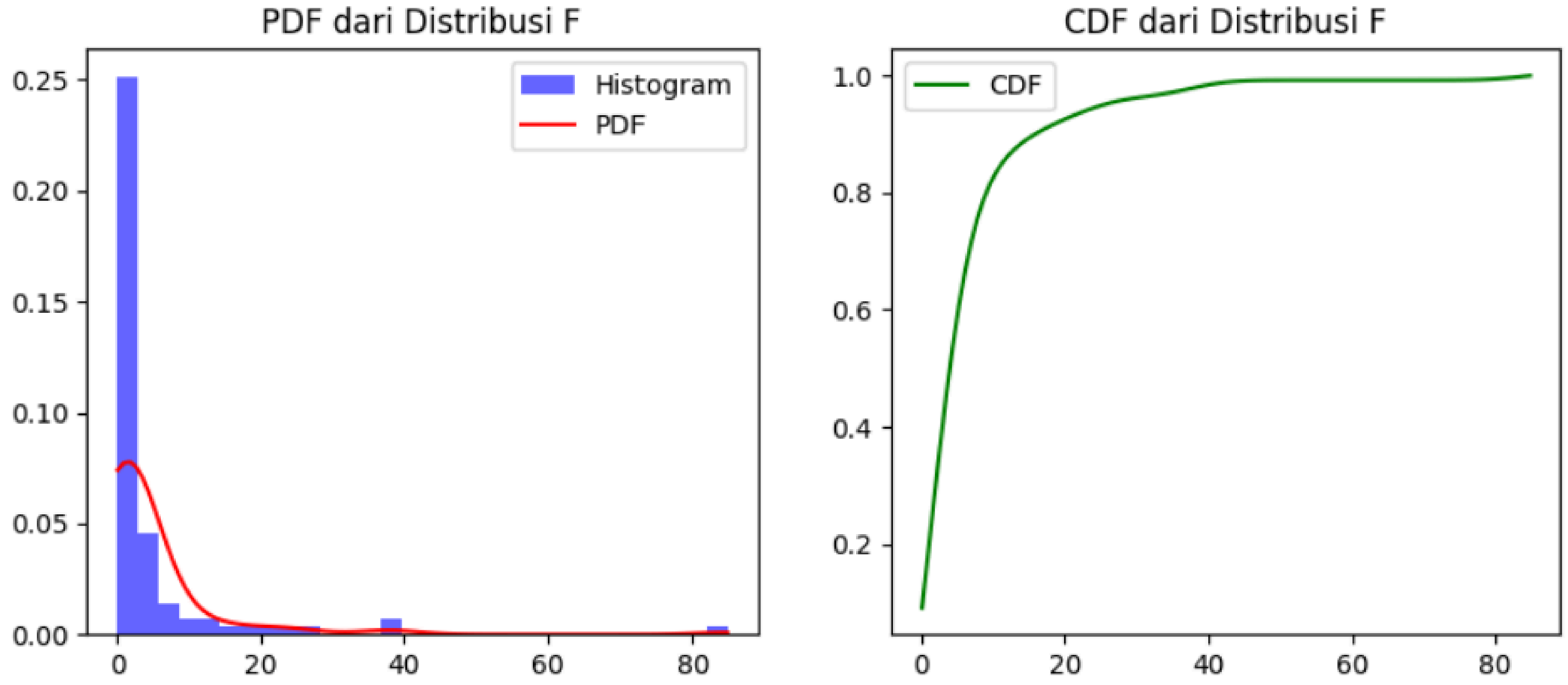
$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t.$$

kode phyton data :

```
distribution = {
    "F": stats.f.rvs(dfn=5, dfd=2, size=100)
}

for name, data in distribution.items():
    x = np.linspace(min(data), max(data), 100)
    pdf = stats.gaussian_kde(data).evaluate(x)  # Estimasi PDF
    cdf = np.cumsum(pdf) / sum(pdf)  # Estimasi CDF

    mean = np.mean(data)
    std = np.std(data)

    print(f"Distribusi {name}: Mean = {mean :.4f}, Std = {std:.4f}")

    plt.figure(figsize=(10, 4))
    plt.subplot(1, 2, 1)
    plt.hist(data, bins=30, density=True, alpha=0.6, color='b', 
label='Histogram')
    plt.plot(x, pdf, color='red', label='PDF')
    plt.title(f"PDF dari Distribusi {name}")
    plt.legend()

    plt.subplot(1, 2, 2)
    plt.plot(x, cdf, color='green', label='CDF')
    plt.title(f"CDF dari Distribusi {name}")
    plt.legend()

    plt.show()
```



visualisasi data :



Distribusi F : Mean = 4.5416, Std = 10.5911

Sains Data

DISTRIBUSI POISSON

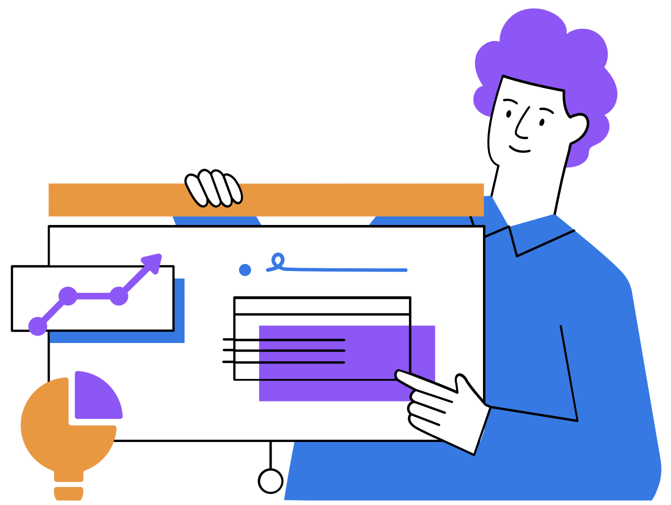
Digunakan untuk menghitung jumlah kejadian dalam interval tetap

pdf :

$$P(x = k) = \frac{\mu^k e^{-\mu}}{k!}$$

cdf :

$$F(x) = \sum_{k=0}^{x} P(X = k)$$

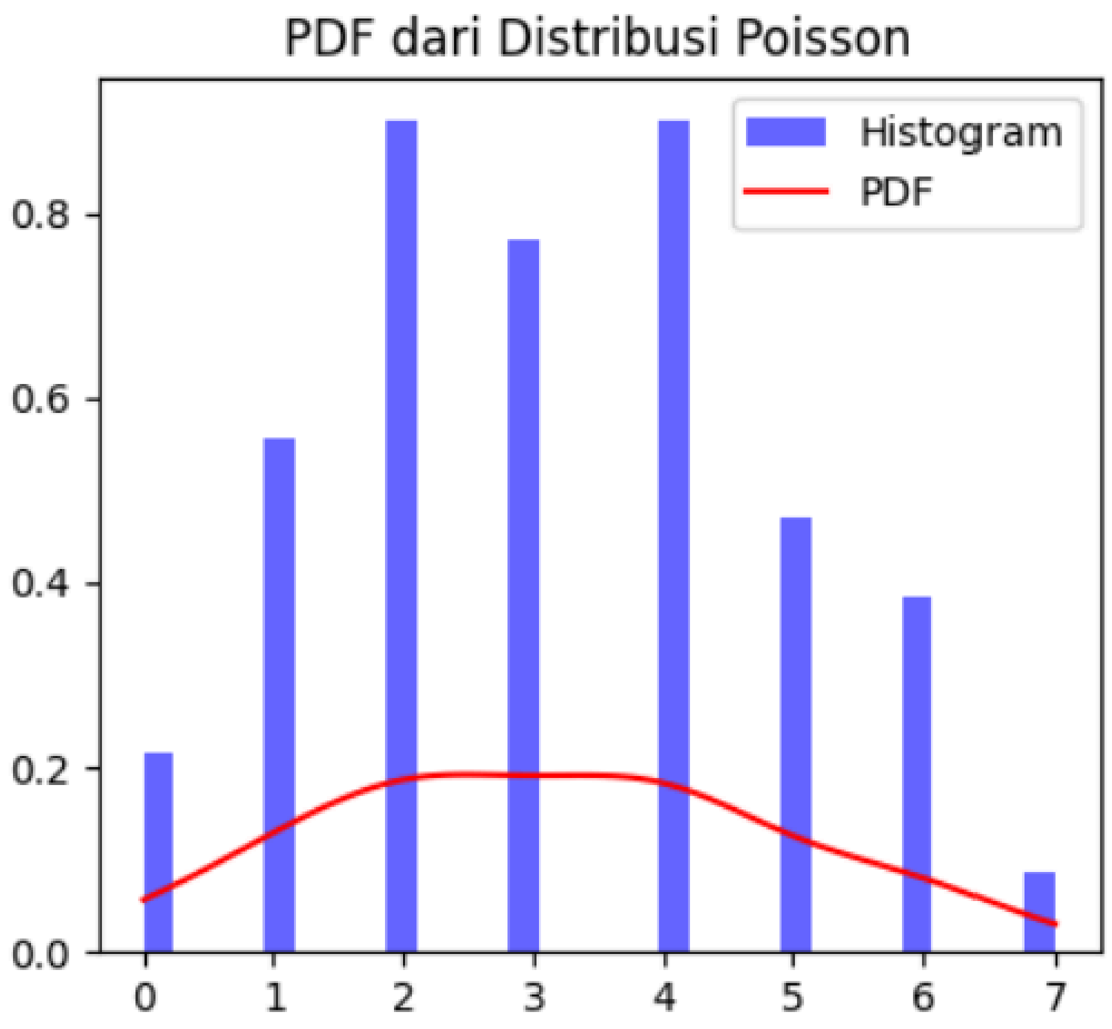


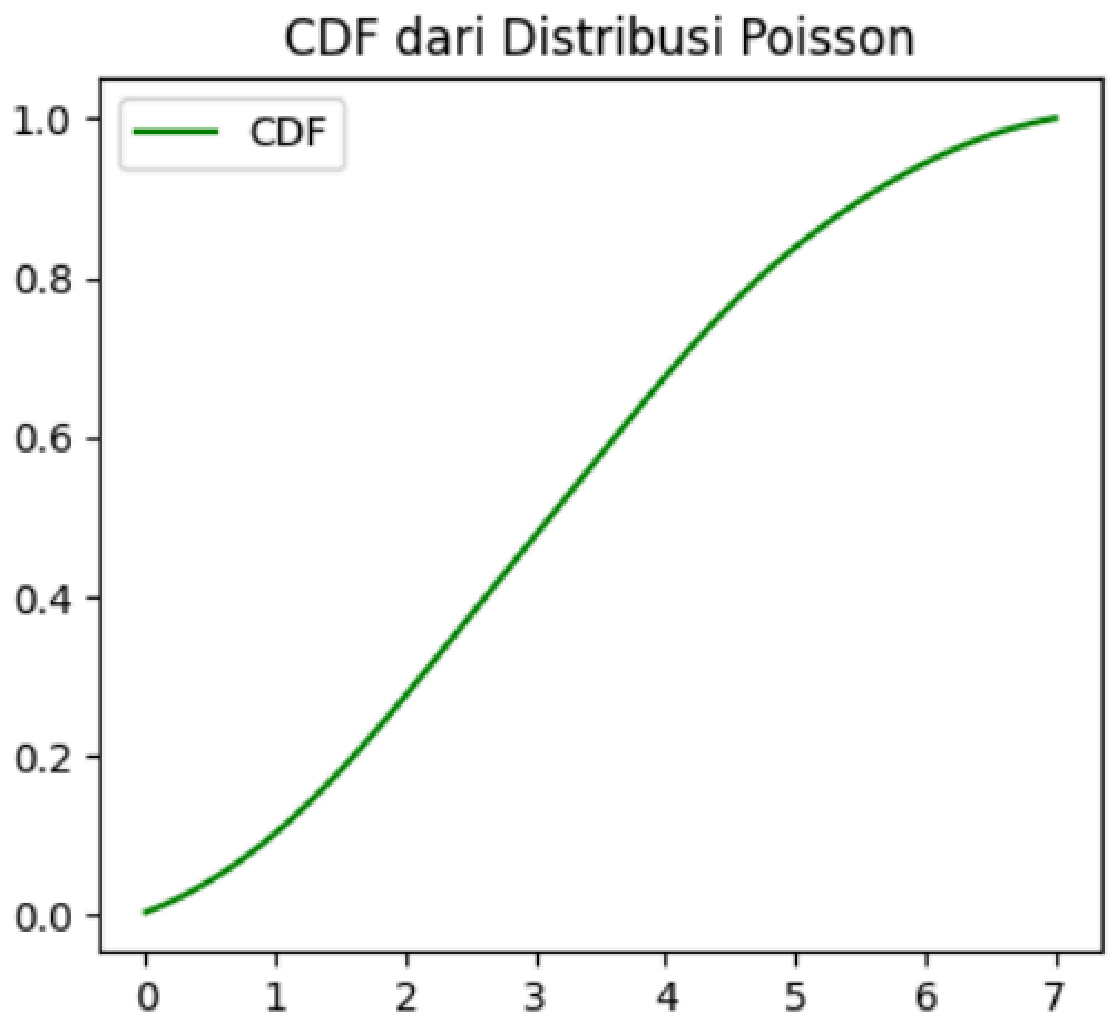
kode phyton data :

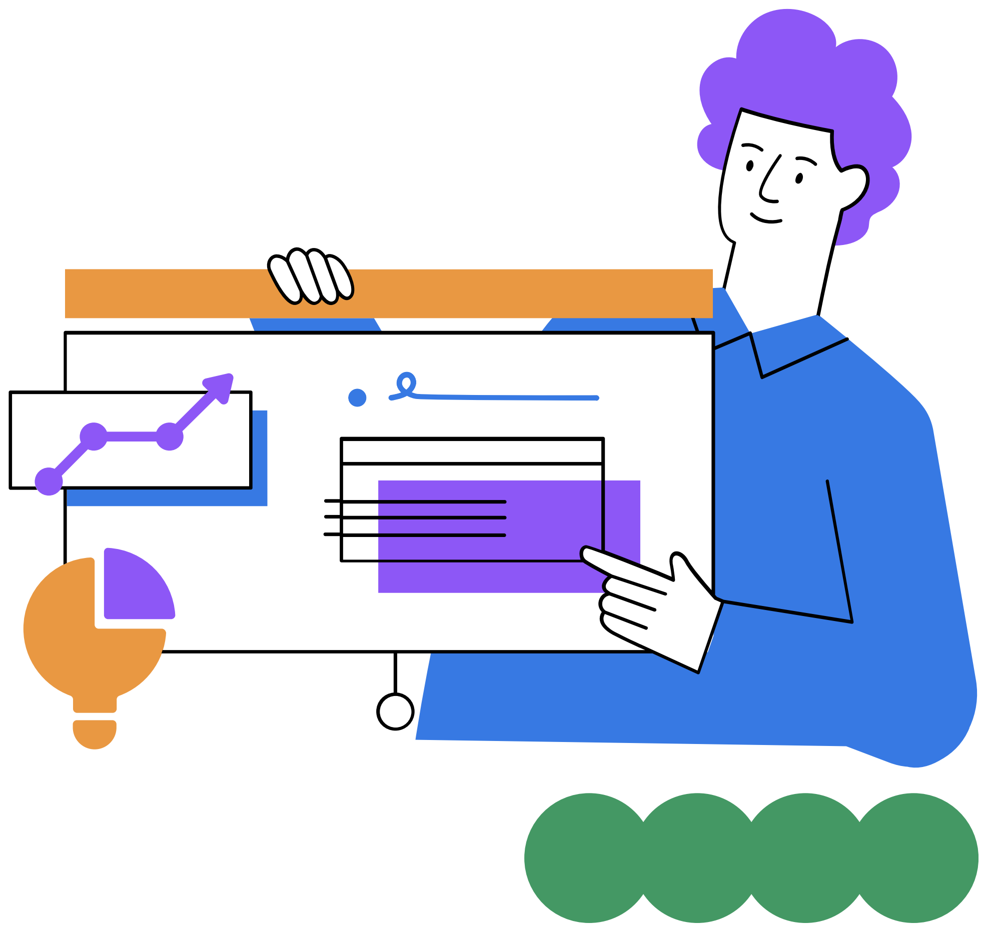
```
distribution = {
    "Poisson": stats.poisson.rvs(mu=3, size=100)
}

for name, data in distribution.items():
    x = np.linspace(min(data), max(data), 100)
    pdf = stats.gaussian_kde(data).evaluate(x)  # Estimasi PDF
    cdf = np.cumsum(pdf) / sum(pdf)  # Estimasi CDF

    mean = np.mean(data)
    std = np.std(data)

    print(f"Distribusi {name}: Mean = {mean :.4f}, Std = {std:.4f}")

    plt.figure(figsize=(10, 4))
    plt.hist(data, bins=30, density=True, alpha=0.6, color='b',
label='Histogram')
    plt.plot(x, pdf, color='red', label='PDF')
    plt.title(f"PDF dari Distribusi {name}")
    plt.legend()

    plt.subplot(1, 2, 2)
    plt.plot(x, cdf, color='green', label='CDF')
    plt.title(f"CDF dari Distribusi {name}")
    plt.legend()

    plt.show()
```

Sains Data

visualisasi data :







Distribusi Poisson : Mean = 3.1600, Std = 1.7130

Sains Data

DISTRIBUSI EKSPONENTIAL

Digunakan untuk model waktu tunggu antar kejadian dalam proses Poisson.

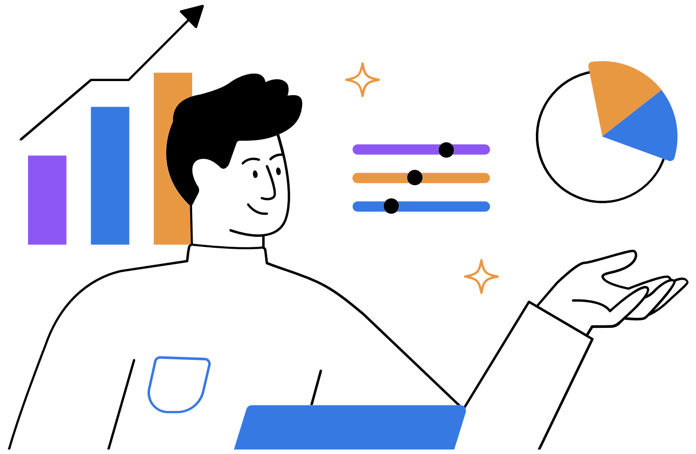
pdf :

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

cdf :

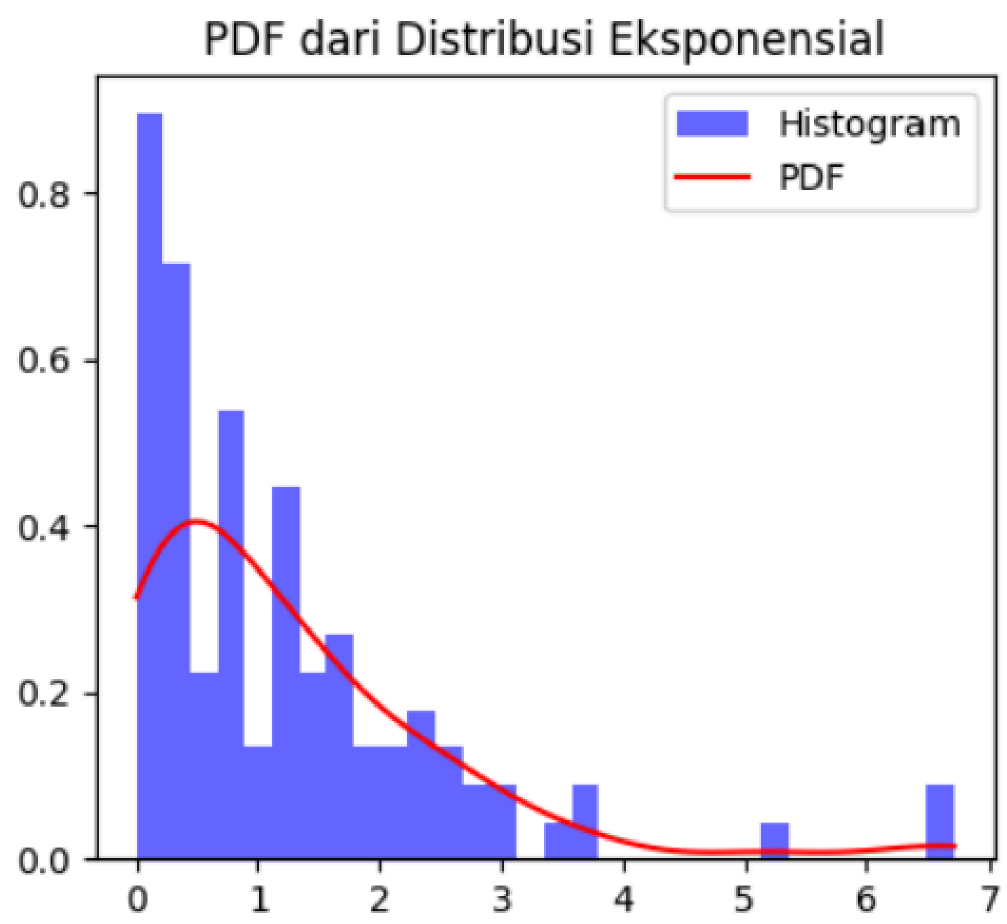
$$F(x) = 1 - e^{-\lambda x}$$

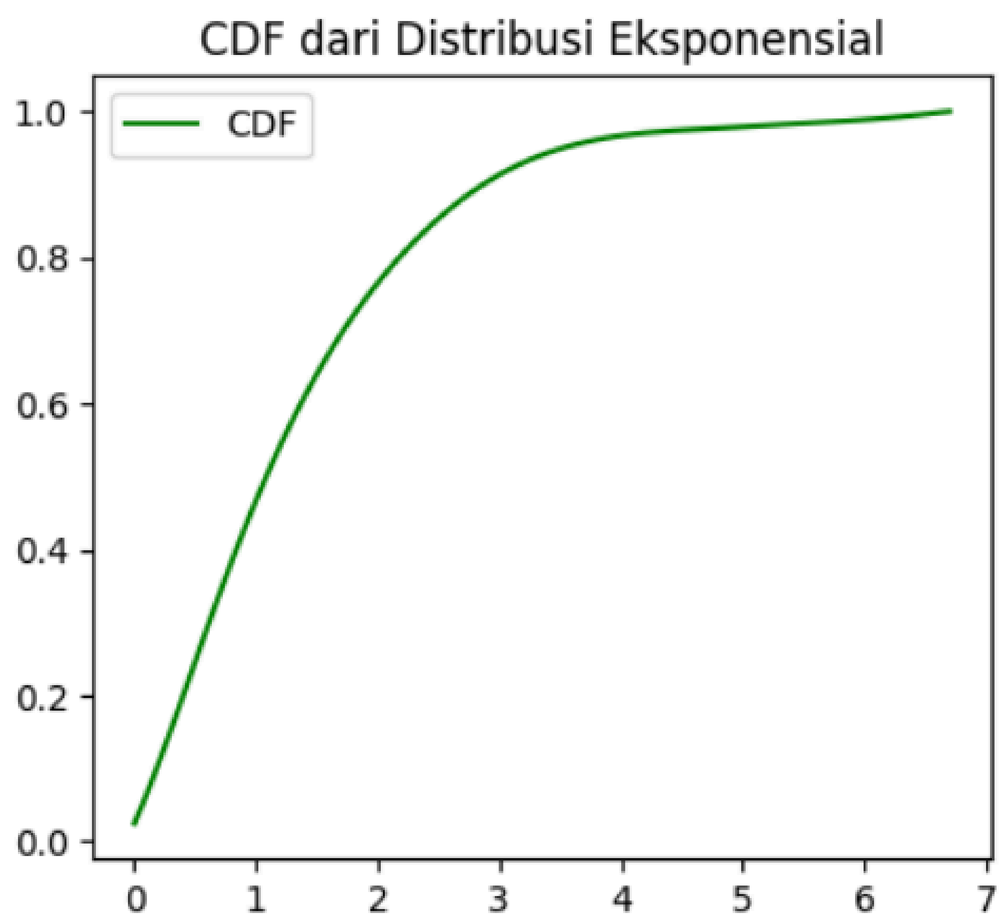
kode phyton data :

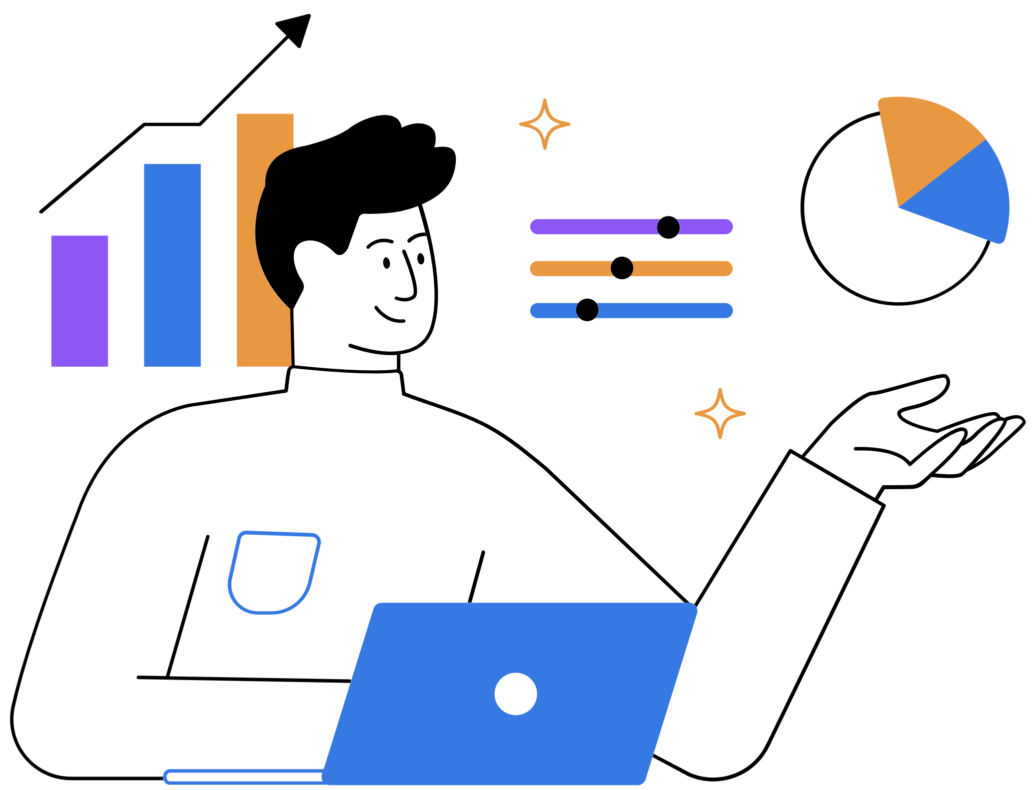


```
distribution = {
    "Eksponensial": stats.expon.rvs(scale=1, size=100)
}

for name, data in distribution.items():
    x = np.linspace(min(data), max(data), 100)
    pdf = stats.gaussian_kde(data).evaluate(x)  # Estimasi PDF
    cdf = np.cumsum(pdf) / sum(pdf)  # Estimasi CDF

    mean = np.mean(data)
    std = np.std(data)

    print(f"Distribusi {name}: Mean = {mean :.4f}, Std = {std:.4f}")

    plt.figure(figsize=(10, 4))
    plt.subplot(1, 2, 1)
    plt.hist(data, bins=30, density=True, alpha=0.6, color='b',
label='Histogram')
    plt.plot(x, pdf, color='red', label='PDF')
    plt.title(f"PDF dari Distribusi {name}")
    plt.legend()

    plt.subplot(1, 2, 2)
    plt.plot(x, cdf, color='green', label='CDF')
    plt.title(f"CDF dari Distribusi {name}")
    plt.legend()

    plt.show()
```

visualisasi data :







Distribusi Eksponential : Mean = 1.2196, Std = 1.2658

Data Sains

DISTRIBUSI WEIBULL

Distribusi fleksibel yang sering digunakan dalam analisis reliabilitas.

pdf :

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, \quad x \geq 0$$

cdf :

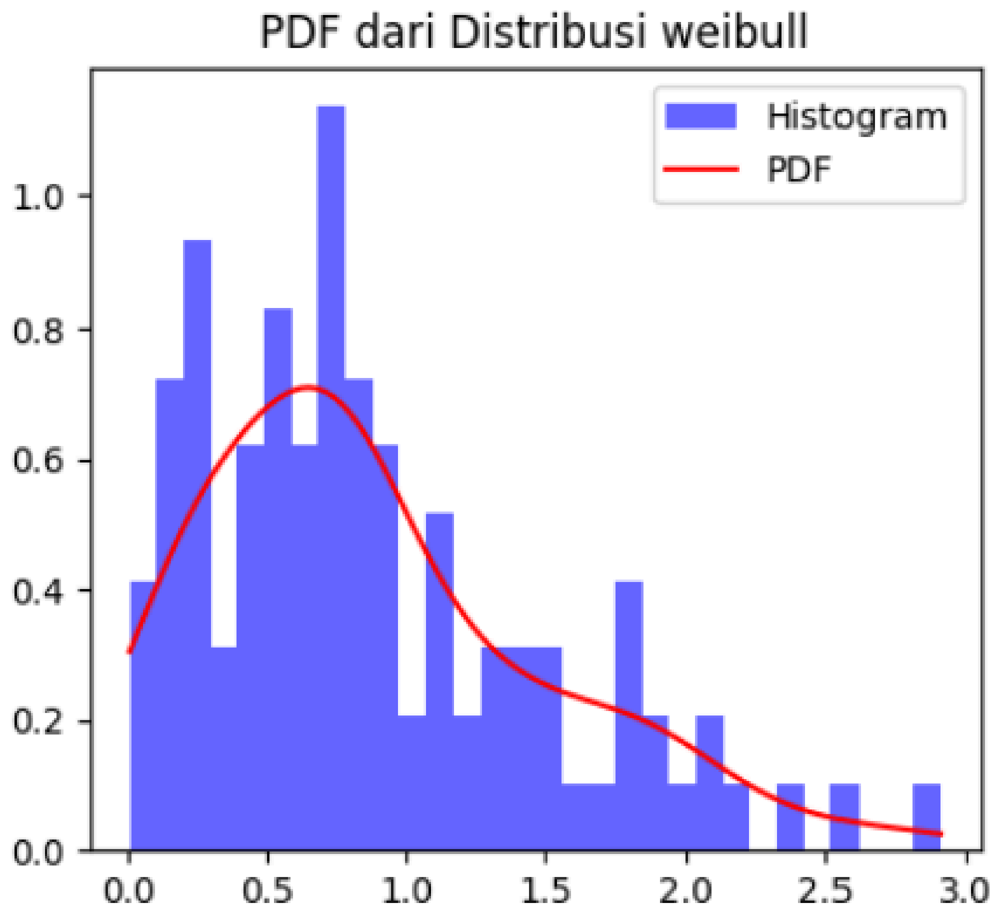
$$F(x) = 1 - e^{-(x/\lambda)^k}$$

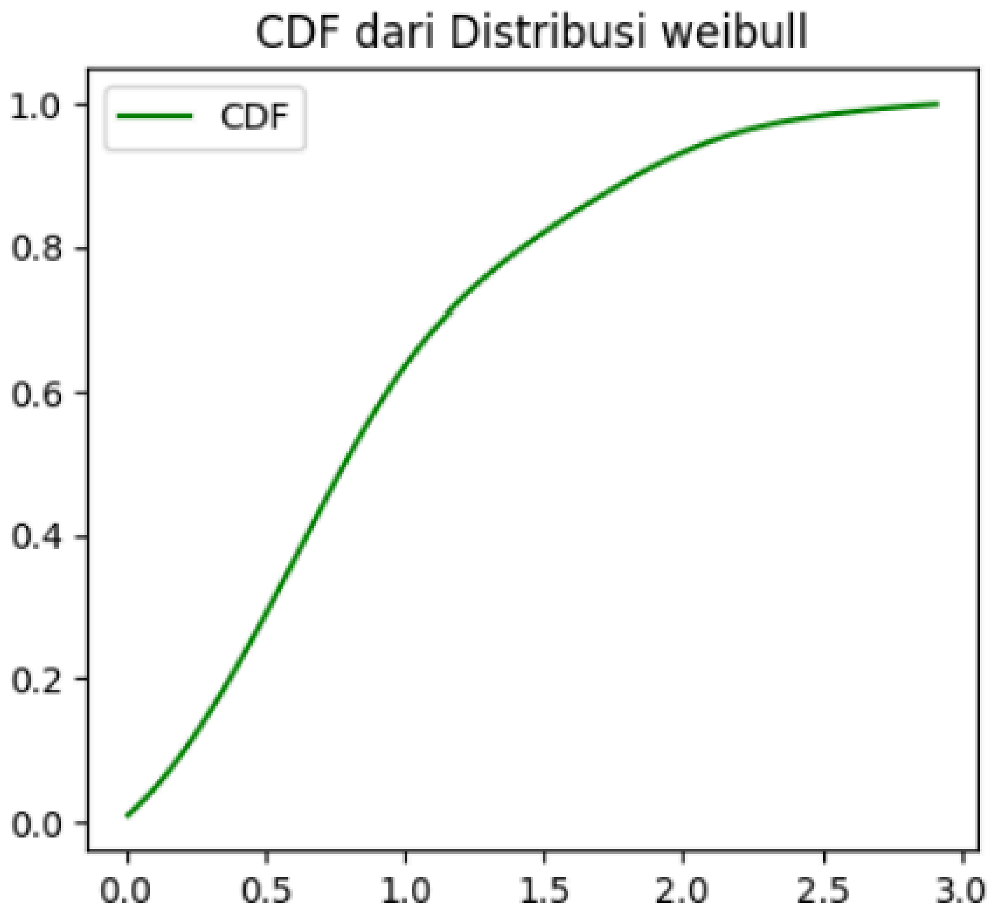
kode phyton data :

```
distribution = {
    "weibull": stats.weibull_min.rvs(c=1.5, size=100)
}

for name, data in distribution.items():
    x = np.linspace(min(data), max(data), 100)
    pdf = stats.gaussian_kde(data).evaluate(x)  # Estimasi PDF
    cdf = np.cumsum(pdf) / sum(pdf)  # Estimasi CDF

    mean = np.mean(data)
    std = np.std(data)

    print(f"Distribusi {name}: Mean = {mean :.4f}, Std = {std:.4f}")

    plt.figure(figsize=(10, 4))
    plt.subplot(1, 2, 1)
    plt.hist(data, bins=30, density=True, alpha=0.6, color='b',
label='Histogram')
    plt.plot(x, pdf, color='red', label='PDF')
    plt.title(f"PDF dari Distribusi {name}")
    plt.legend()

    plt.subplot(1, 2, 2)
    plt.plot(x, cdf, color='green', label='CDF')
    plt.title(f"CDF dari Distribusi {name}")
    plt.legend()
```

Data Sains

Visualisasi data :





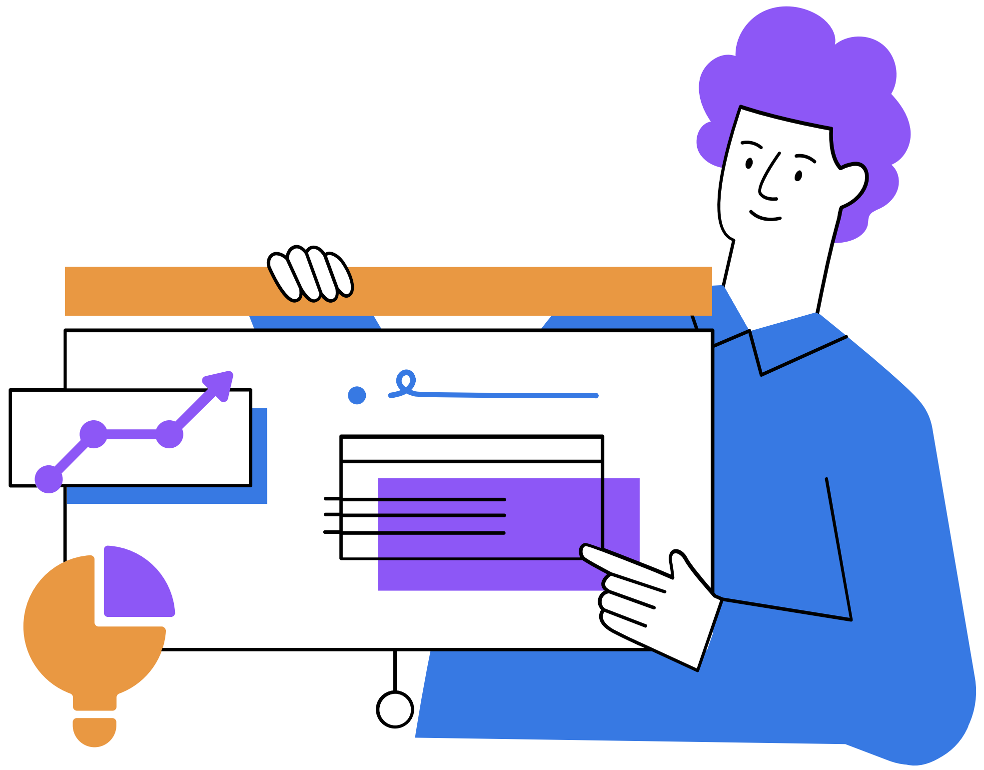
Distribusi Weibull : Mean = 0.8798, Std = 0.6189

tabel perbandingan :

No.	Distribusi	Mean	Standar deviasi
1	Normal	0,0049	1,066
2	Long-tailed	1,7525	17,0255
3	Student's T	0,2042	1,1656
4	Binomial	4,98	1,3265
5	Chi-square	4,4182	2,9563
6	F	4,516	10,5911
7	Poisson	3,16	1,731
8	Eksponensial	1,2196	1,2658
9	Weibull	0,8798	0,6189

kesimpulan :

hasil pembangkitan dan analisis data, setiap distribusi memiliki karakteristik unik yang tercermin dalam bentuk histogram, PDF, dan CDF. Distribusi seperti Normal dan Binomial menunjukkan pola yang lebih terpusat, sementara distribusi seperti long-tailed memiliki ekor panjang yang menyebabkan nilai-nilai ekstrem. Nilai mean dan standar deviasi dari data yang dibangkitkan menunjukkan adanya variasi yang bergantung pada jenis distribusi.



Borcelle

Thank You