SÜLEYMAN DEMİREL ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
VERİ MADENCİLİĞİ VİZE CEVAP ANAHTARI

ADI SOYADI:…….NO:……

Sınav Yönergesi:

1. Sınav Süresi 50 dakikadır
2. Soruların puanları üzerinde yazdığı gibidir.
3. Defter Kitap açıktır. Her soru altındaki boşluğa cevaplanacaktır.
4. Cep telefonunun yanınızda veya yakınınızda bulunması kesinlikle yasaktır.

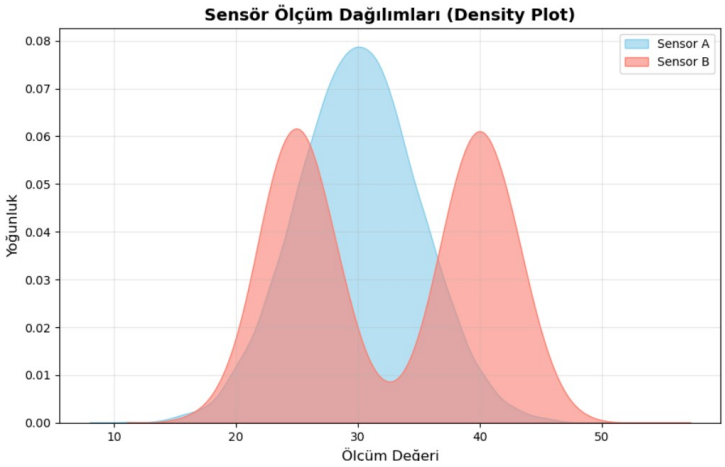
1. Bir üretim tesisinde üretilen çiplerin hata sayısı, geçmiş verilere göre ortalama $\lambda = 3$ ile Poisson dağılımına uymaktadır. Yeni üretim hattı kurulduktan sonra, 1000 çip üzerinde test yapılmış ve her bir çipteki hata sayısı kayıt altına alınmıştır.**(30p)**
a) Poisson(λ=3) dağılımına göre 1000 gözlem üretiniz. b) Bu verilerin ortalama ve varyansını hesaplayınız, c) Histogramını çiziniz ve üzerine teorik Poisson PMF eğrisini ekleyiniz. d) Yeni üretim hattında hata sayısının 5'ten büyük olma olasılığını hesaplayınız?
Yukarıdaki işlemleri yapan python kodunu yazınız?

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

# (a) 1000 gözlem üret
np.random.seed(42)
data = poisson.rvs(mu=3, size=1000)

# (b) Ortalama ve varyans
mean_emp = np.mean(data)
var_emp = np.var(data)

# (c) Histogram + teorik PMF
x = np.arange(0, max(data) + 1)
plt.hist(data, bins=x, density=True, alpha=0.6, color="skyblue", label="Simülasyon")
plt.plot(x, poisson.pmf(x, mu=3), 'r--', lw=2, label="Teorik PMF")
plt.title("Poisson(λ=3) - Simülasyon vs Teorik Dağılım")
plt.xlabel("Hata Sayısı")
plt.ylabel("Olasılık Yoğunluğu")
plt.legend()
plt.show()

p_analitik = 1 - poisson.cdf(5, mu=3)
```

2. İçinde bulunulan klasördeki tüm .csv dosyalarının içinde "temperature" kelimesinin geçtiği satır sayısı bulunmak istenmektedir. Bu işlemi tek satırda yapacak Linux terminal komutunu yazınız?**(10p)**

```
grep -i "temperature" *.csv | wc -l
```

3. Aşağıda iki farklı sensörün (Sensor A ve Sensor B) veri dağılımları aynı ölçekte çizilmiştir. Grafik, her sensörden toplanan 10.000 ölçümün kernel density estimation (KDE) grafiğidir.**(20p)**



a) Her iki sensörün dağılım şekillerini tanımlayınız?

Sensor A: Tek modlu (unimodal), simetrik dağılım. Normal dağılıma oldukça yakın.
Sensor B: Çift modlu (bimodal) yapı: iki ayrı tepe noktası.

b)Hangi sensörde veri varyansı daha yüksektir? Neden?

Sensor B'nin varyansı daha yüksek. Çünkü veri iki farklı moda dağılmış, değer aralığı daha geniş.

c) Sensor A'nın verilerinin normal dağılıma yakın olup olmadığını hangi ölçütlerle sınayabilirsiniz?

Çapıklık ve basıklık

4. Bir e-ticaret firması, yeni bir ürün sayfası tasarımının kullanıcıların site üzerinde ortalama kalma süresini artıracağı iddiasında bulunuyor. Eski sayfada kullanıcıların ortalama kalma süresi 4.2 dakikadır. Yeni tasarım 50 kullanıcı üzerinde test edilmiş, örneklem ortalaması 4.5 dakika, standart sapma 1.0 dakika olarak ölçülmüştür. Firma yöneticisi, "yeni tasarım ortalama kalma süresini artırdı" diyerek bunu istatistiksel olarak kanıtlamak istemektedir. Buna göre;**(20p)**

a) H_o ve H_a hipotezlerini yazınız?

H_o: μ=4.2(Yeni tasarımın ortalama etkisi yok)
H_a: μ>4.2(Yeni tasarım ortalamayı artırdı)

b) Hangi kuyruk testi yapılmalıdır?

Sağ Kuyruk testi yapılmalıdır.

5. Aşağıdaki **Pandas** ifadelerini tek bir SQL sorgusu olarak ifade ediniz **? (20p)**

```
import pandas as pd
# Varsayalım DataFrame adı df
filtered = df[df["passed_bool"] == True]
grouped = (
   filtered.groupby("department", as_index=False)
        .agg(student_count=("student_id", "count"),
            avg_score=("exam_score", "mean"))
)
result = grouped[grouped["avg_score"] > 70]
```

SELECT department, COUNT(*) AS student_count, AVG(exam_score) AS avg_score
FROM students WHERE passed_bool = TRUE GROUP BY department
HAVING AVG(exam_score) > 70;