DSA 8301 - Statistical Inference in Big Data
CAT # 1 (due 11:59 PM EAT May 21, 2025)

INSTRUCTIONS:

1. **You must show *ALL* work to receive *ANY CREDIT*.**
2. **Submit your own work. Do not consult *ANYONE*. Violations will be heavily penalized.**

1) Show that 63/512 is the probability that the fifth head is observed on the tenth independent flip of a fair coin.

2) Let X be distributed as $\Gamma(\alpha, \beta)$. Find

a) the method of moment estimators of α and β.

b) the point estimates of α and β based on the method of moments and the data below:

```
6.9 7.3 6.7 6.4 6.3 5.9 7.0 7.1 6.5 7.6 7.2 7.1 6.1
7.3 7.6 7.6 6.7 6.3 5.7 6.7 7.5 5.3 5.4 7.4 6.9
```

3) Let $X_1, ..., X_{10}$ be a random sample from a population with mean μ and variance σ^2, and $Y_1, ..., Y_{10}$ be a random sample from another population with mean also equal to μ and variance $4\sigma^2$. The two samples are independent.

a) Show that for any α, $0 \leq \alpha \leq 1$, $\hat{\mu} = \alpha\overline{X} + (1-\alpha)\overline{Y}$ is a unbiased estimator for μ.

b) Obtain an expression for the MSE of $\hat{\mu}$.

c) Is the estimator $\overline{X}$ preferable over the estimator $0.5\overline{X} + 0.5\overline{Y}$? Justify your answer.

4) Let $X_1, ..., X_n$ be a random sample from $\mathrm{N}(\mu, \sigma^2)$, where σ^2 is known.

a) Show that $Y = (X_1 + X_2)/2$ is an unbiased estimator of μ.

b) Find the Cramer-Rao lower bound for the variance of an unbiased estimator of μ for a general n.

c) What is the efficiency of Y in part (a) above?

5) If $X_1, X_2, ..., X_n$ is a random sample from a distribution having $p.d.f$ of the form

$$f(x;\theta) = \theta x^{\theta-1},\ 0 < x < 1,$$

show that a best critical region for testing $H_0 : \theta = 1$ vs $H_a : \theta = 2$ is $C = ((x_1, x_2, ..., x_n) : c \leq \prod_{i=1}^{n} x_i)$.

6) Let $X_1, X_2, ..., X_n$ be *i.i.d.* $\mathrm{N}(\theta_1, \theta_2)$. Show that the likelihood ratio principle for testing $H_0 : \theta_2 = \theta_2^{'}$ specified, and θ_1 unspecified vs. $H_a : \theta_2 \neq \theta_2^{'}$, θ_1 unspecified, leads to a test that rejects when

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 \leq c_1 \text{ or } \sum_{i=1}^{n}(x_i - \overline{x})^2 \geq c_2,$$

where $c_1 < c_2$ are selected appropriately.

7) Let X have the $p.m.f.$

$$f(x;\theta) = \theta^x(1-\theta)^{1-x}, x = 0, 1.$$

We test $H_0 : \theta = 0.5$ vs. $H_a : \theta < 0.5$ by taking a random sample $X_1, X_2, ..., X_n$ of size $n = 5$ and rejecting H_0 if $Y = \sum_{i=1}^{n} X_i$ is observed to be less than or equal to a constant c.

a) Show that this is a uniformly most powerful test.
b) Find the significance level when $c = 1$.
c) Find the significance level when $c = 0$.

8) In a biology laboratory, students use corn to test the Mendelian theory of inheritance. The theory claims that frequencies of the four categories "smooth and yellow", "wrinkled and yellow", "smooth and purple", and "wrinkled and purple" will occur in the ratio $9 : 3 : 3 : 1$. If a student counted 124, 30, 43, and 11, respectively, for these four categories, would these data support the Mendelian theory? Let $\alpha = 0.05$.