DSA 8301 - Statistical Inference in Big Data
CAT # 1 (due 11:59 PM EAT June 11, 2023)

You must show *all* work to receive *any credit*.

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 MLE of α and β.

c) 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) Researchers studying healthy body composition recorded various measurements of 72 male subjects. Height, weight, neck, age, and other measurements were collected on each subject and stored in the data set "***bodymeasgrps.txt***". They want to predict the height of a man given his neck length. Use this information to answer questions (a) through (e) below:

a) If we consider the relationship between the height and neck length of these men, which variable should go on the X-axis? The Y-axis? Justify your answer.

b) Create and attach a scatterplot of these two variables. Include the linear regression line on your plot.

c) Find the equation of the linear regression line used to predict the height of a man given his neck length.

d) Interpret the slope and intercept of the regression line, if appropriate. If it is not appropriate to make an interpretation, explain why not.

e) Predict the height of a man whose neck is 35 inches long in two ways: using the equation for the regression line AND using R.

4) Facilities A and B account for 40% and 60%, respectively, of the production of a certain electronic component. The two components from the two facilities are shipped to a packaging location where they are mixed and packaged. A sample of size 100 will be used to estimate the expected life time in the combined population. Use the MSE criterion to decide which of the following two sampling schemes (simple random versus stratified sampling) should be adapted, i.e. simple random sampling at the packaging location, and stratified random sampling based on a simple random sample of size 40 from facility A and a simple random sample of size 60 from facility B. (**Hint**: $\mu = 0.4\mu_A + 0.6\mu_B$, $\sigma^2 = 0.4\sigma_A^2 + 0.6\sigma_B^2 + (0.6)(0.4)(\mu_B - \mu_A)^2$ and the estimators of the mean under both sampling schemes are unbiased for μ.)

5) Let $X_1, ..., X_{10}$ be a random sample from a population with mean mu 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, $0 \le \alpha \le 1$, $\hat{\mu} = \alpha\overline{X} + (1-\alpha)\overline{Y}$ 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.

6) 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?