NTHU MATH 2820 Midterm Examination May 19, 2020

Note. There are 6 problems in total. **For problems 3 to 6, to ensure consideration for partial scores, write down necessary intermediate steps.** Correct answers with inadequate or no intermediate steps **may result in zero credit**.

Some useful formula.

- Suppose that X follows a binomial(n, p) distribution. The probability mass function (pmf) of X is
$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad \text{for} \;\; x = 0, 1, \ldots, n.$$
The mean of X is np, and the variance of X is $np(1-p)$.

- Suppose that $(X_1, \ldots, X_m)$ follows a multinomial$(n, m, p_1, \ldots, p_m)$ distribution, where $p_1 + \cdots + p_m = 1$. The joint pmf of $(X_1, \ldots, X_m)$ is
$$p(x) = \binom{n}{x_1 \cdots x_m} p_1^{x_1} \cdots p_m^{x_m},$$
for $x_i = 0, \ldots, n$, where $i = 1, \ldots, m$, and $x_1 + \cdots + x_m = n$. For $i = 1, \ldots, m$, the marginal distribution of X_i is binomial(n, p_i). For $1 \leq i < j \leq m$, $\text{cov}(X_i, X_j) = -np_ip_j$.

- Suppose that X follows a uniform(a, b) distribution. The probability density function (pdf) of X is
$$f(x) = \frac{1}{b-a}, \quad \text{for} \;\; a < x < b,$$
and zero otherwise. The cumulative distribution function (cdf) of X is
$$F(x) = \frac{x-a}{b-a}, \quad \text{for} \;\; a < x < b,$$
and zero if $x \leq a$ and one if $x \geq b$. The mean of X is $\frac{a+b}{2}$, and the variance of X is $\frac{(b-a)^2}{12}$.

- Suppose that X follows an exponential(λ) distribution. The pdf of X is
$$f(x) = \lambda e^{-\lambda x}, \quad \text{for} \;\; 0 < x < \infty,$$
and zero otherwise. The cdf of X is
$$F(x) = 1 - e^{-\lambda x}, \quad \text{for} \;\; 0 < x < \infty,$$
and zero otherwise. The mean of X is $1/\lambda$, and the variance of X is $1/\lambda^2$.

---

1. (*14 pts in total. for each question,* **2 pts if correct; 0 pt if blank; −1 pt if wrong**) For the following statements, please answer true or false.

(a) Convergence almost surely implies convergence in probability, and convergence in probability implies convergence in distribution.

(b) A point *estimate* (i.e. an estimated value) of a parameter θ is a random variable with an associated probability distribution, called sampling distribution.

(c) Because sufficient statistics contain *all* information about the unknown parameters, we can *always* throw away all original data and keep only the sufficient statistics.

(d) In the asymptotic method, the exact sampling distribution is often replaced by a normal distribution while in the (parametric) bootstrap method, the estimated value of the parameter is often treated as the true parameter.

(e) No estimators can have a smaller variance than the UMVUE.

~~(f) In general, the width of a confidence interval would increase when the sample size is doubled.~~

~~(g) A pivotal quantity is a random variable, but not a statistic.~~

2. (*12pts, 4pts for each*) For each of the random observation(s) $\boldsymbol{X}$ given below,

- determine the type of the *joint distribution* (e.g., Normal, Exponential, Gamma, Beta, Uniform, Poisson, Hyper-geometric, Bernoulli, Binomial, Multinomial, Negative binomial, Geometric, or others) that best models $\boldsymbol{X}$, and
- identify the known and unknown parameter(s) in the chosen distribution.

(a) A company had manufactured certain objects and had printed a serial number on each manufactured object. The serial numbers start at 1 and end at the total number of objects that had been manufactured. The record about how many objects had been manufactured by the company was lost. To understand the total number, we randomly selected n objects *with replacement*. Let $X_1, \ldots, X_n$ be the serial numbers of the selected objects. Set $\boldsymbol{X} = (X_1, \ldots, X_n)$.

(b) George and Hilary wanted to investigate the probability of obtaining a head when a coin is tossed. George tossed the coin three times. Let X_1 be the number of heads that George observed. He then gave the coin to Hilary. She tossed it until the first head occurs. Let X_2 be the number of times that Hilary tossed the coin. Set $\boldsymbol{X} = (X_1, X_2)$.

(c) A sample of 311 patients were randomly chosen from a population consisting of the cigarette-smoking patients of residents in family medicine seen in a medical center. The 311 patients were cross-classified into four categories formed by two factors (see the table given below): (i) a patient was advised or was not advised to stop smoking by the physicians; (ii) race of the patient (White and other versus African-American).

factor (ii) (race)	factor (i) Advised	Not Advised
White and Other	$X_{W,1}$	$X_{W,2}$
African-American	$X_{A,1}$	$X_{A,2}$

The numbers of the patients in each categories were denoted by $X_{W,1}$, $X_{W,2}$, $X_{A,1}$, $X_{A,2}$, respectively. It was known that whether a patient was advised by the physicians or not is *independent* of the race of the patient. Set $\boldsymbol{X}$=($X_{W,1}$, $X_{W,2}$, $X_{A,1}$, $X_{A,2}$).

3. Let $Y_1, \ldots, Y_n$ be i.i.d. from the uniform$(\theta, \theta+1)$ distribution.

   (a) (*3pts*) Show that the method of moments estimator for θ (denoted as $\hat{\theta}_1$) is
$$\overline{Y} - \frac{1}{2},$$
   and show that $\hat{\theta}_1$ is an unbiased estimator of θ.

   (b) (*2pts*) Find the standard error of $\hat{\theta}_1$.

   (c) (*4pts*) Let
$$\hat{\theta}_2 = Y_{(n)} - \frac{n}{n+1},$$
   where $Y_{(n)} = \max\{Y_1, \ldots, Y_n\}$. Show that $\hat{\theta}_2$ is an unbiased estimator of θ. [**Hint. (i)** Let $T_i = Y_i - \theta$, then T_i has a uniform(0, 1) distribution. **(ii)** Let $T_{(n)} = \max\{T_1, \ldots, T_n\}$, then $T_{(n)} = Y_{(n)} - \theta$. Therefore, $\mathrm{E}(T_{(n)}) = \mathrm{E}(Y_{(n)} - \theta)$ and $\mathrm{Var}(T_{(n)}) = \mathrm{Var}(Y_{(n)})$.]

   (d) (*4pts*) Find the standard error of $\hat{\theta}_2$.

   (e) (*2pts*) Show that the efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$, denoted as $\mathrm{eff}_n(\hat{\theta}_1, \hat{\theta}_2)$, is
$$\frac{12n^2}{(n+1)^2(n+2)},$$
   and find their asymptotic relative efficiency.

   (f) (*3pts*) When the sample size n is greater than 7, explain which of $\hat{\theta}_1$ and $\hat{\theta}_2$ is better in terms of *mean square error*. [**Hint.** $\mathrm{eff}_n(\hat{\theta}_1, \hat{\theta}_2) < 1$, when $n > 7$.]

4. Let $Y_1, \ldots, Y_n$ be i.i.d. from the pdf:
$$f(y|\theta) = \begin{cases} \left(\frac{1}{\theta}\right) r y^{r-1} e^{-\frac{y^r}{\theta}}, & y > 0 \\ 0, & \text{otherwise} \end{cases}$$
   where r is a known positive constant and θ is a positive parameter (i.e., $\theta > 0$).

   (a) (*4pts*) Use the factorization theorem to find a sufficient statistic.

   (b) (*4pts*) Find the maximum likelihood estimator (MLE) of θ.

   (c) (*4pts*) Show that Y_1^r has an exponential$(\frac{1}{\theta})$ distribution.

   (d) (*5pts*) Calculate Fisher information contained in $Y_1, \ldots, Y_n$, and from which find the asymptotic variance of the MLE. [**Hint.** Use 4(c).]

5. Suppose that in the population of twins, male (M) and females (F) are equally likely to occur. Suppose that the probability that twins are identical (i.e., they result from the fertilization of a *single egg* that splits in two) is α, where α is an unknown constant. If twins are not identical, their genes are independent. Use MM to denote the cases that twins are male, FF to denote the cases that twins are female, and MF to denote the cases that twins are one male and one female.

(a) (*4pts*) Show that

$$P(MM) = P(FF) = \frac{1+\alpha}{4}, \quad \text{and} \quad P(MF) = \frac{1-\alpha}{2}.$$

[**Hint.** First of all, find the conditional probabilities $P(z|\text{a twin is identical})$ and $P(z|\text{a twin is not identical})$, where $z = MM, FF$, or MF.]

(b) (*2pts*) Suppose that n (a known number) twins are randomly sampled. It is found that X_1 twins of them are MM, X_2 twins are FF, and X_3 twins are MF, but it is not known which twins are identical. Use 5(a) to identify the joint distribution of (X_1, X_2, X_3). [**Hint.** Use multinomial distribution.]

(c) (*4pts*) Show that the joint pmf of (X_1, X_2, X_3) forms a one-parameter exponential family, and find a sufficient and complete statistic for α. [**Hint.** Replace X_3 by $n - X_1 - X_2$.]

(d) (*4pts*) Show that the MLE of α is

$$\hat{\alpha} = \frac{X_1 + X_2 - X_3}{n}.$$

[**Hint.** $X_1 + X_2 + X_3 = n$.]

(e) (*2pts*) Show that $\hat{\alpha}$ is an unbiased estimator.

(f) (*4pts*) Find the exact variance of $\hat{\alpha}$.

(g) (*5pts*) Show that the Cramer-Rao lower bound for all unbiased estimators of α is

$$\frac{1-\alpha^2}{n}.$$

(h) (*2pts*) Explain why $\hat{\alpha}$ has the smallest variance among all unbiased estimators of α.

6. ~~Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample with the cdf:~~

$$F(x|\theta) = \begin{cases} 1 - e^{-(x-\theta)}, & \theta < x < \infty \\ 0, & \text{otherwise} \end{cases},$$

~~where $\theta \in (-\infty, \infty)$. Let $Y = X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$, then:~~

~~(a) (*2pts*) Show that the cdf of Y is~~

$$F_Y(y|\theta) = \begin{cases} 1 - e^{-n(y-\theta)}, & \theta < y < \infty \\ 0, & \text{otherwise} \end{cases}.$$

~~[**Hint**: $P(Y \le y) = 1 - P(Y > y) = 1 - P(X_1 > y, X_2 > y, \ldots, X_n > y)$.]~~

~~(b) (*4pts*) Show that $T_n(\theta) = 2n(Y - \theta)$ is a pivotal quantity. [**Hint**: use 6(a) to find the cdf of $T_n(\theta)$.]~~

~~(c) (*6pts*) Use the pivotal quantity in (b) to find a $100(1-\alpha)\%$ confidence interval for θ. **Please express the confidence interval in terms of Y, n, and α.**~~