This is a take home midterm. You can use your notes, my online notes on canvas and the textbooks book. You are supposed to work on your own text without external help. I'll be available to answer question in person or via email. Please, write clealy and legibly and take a readable scan before uploading.

Name (print): ________________________________

Question:	1	2	3	Total
Points:	45	35	20	100
Score:				

Question:	1	2	3	Total
Bonus Points:	20	0	0	20
Score:				

Question 1 .... *45 point*

Parameter free statistics: Let X_i, $i = 1, \dots, N$ be a random sample where each X_i is a continuous r.v. with p.d.f. $f(x)$ (and c.d.f $F(x)$). Let Y_i be the order statistics, that is $Y_1 = \min\{X_i\}$ is the smallest among the X_i, $Y_k = \min\{X_i \,|\, X_i > Y_{k-1}\}$ is the k-th smallest among the X_i, till $Y_N = \max\{X_1\}$. Observe that the Y_i are statistics, that is, they function of the random variables forming the sample.

(a) (10 points) Show that

$$\mathbb{P}(Y_k \leq y) = \sum_{j=k}^{N} \binom{N}{j} F(y)^j (1 - F(y))^{N-j}\,.$$

(**Hint**: the N events $\{X_i \leq y\}$ are independent.)

Solution: If $Y_k < y$, at least k among the X_i are smaller than y. Since the X_i are independent and each has a probability probability $F(y)$ to be smaller than y it follows that the number of X_i less than y is a binomial r.v. with parameters N and $F(y)$.

(b) (10 points) Let m be the median of the population, that is m satisfies

$$F(m) = \frac{1}{2}\,.$$

For $i < N/2$, find

$$\mathbb{P}(Y_i \leq m \leq Y_{N-i+1})\,.$$

Solution: We have

$$\begin{aligned}\mathbb{P}(Y_i \leq m \leq Y_{N-i+1}) =& \mathbb{P}(m \leq Y_{N-i+1}) - \mathbb{P}(m < Y_i) = \\ & \mathbb{P}(Y_i \leq m) - \mathbb{P}(Y_{N-i+1} < m) = 2^{-N} \sum_{j=i}^{N-i} \binom{N}{j} = \\ & \mathrm{Bin}(N - i, N, 0.5) - \mathrm{Bin}(i - 1, N, 0.5) = \\ & 1 - 2\mathrm{Bin}(i - 1, N, 0.5)\end{aligned}$$

where $\mathrm{Bin}(x; N, p)$ is the c.d.f. of a binomial r.v. with parameters N and p.

(c) (10 points) Assume now $N = 20$. Use point b) to find a coefficient 0.95 confidence interval for the median m. You can use any software you want to do the computation. This is an online binomial calculator.

Solution: Observing that

$$\text{Bin}(5, N, 0.5) = 2^{-20} \sum_{j=0}^{5} \binom{20}{j} = 0.020695$$

while

$$\text{Bin}(6, N, 0.5) = 2^{-20} \sum_{j=0}^{6} \binom{20}{j} = 0.057659$$

we get

$$\mathbb{P}(Y_6 \leq m \leq Y_{15}) = 0.95861$$

so that

$$Y_6 \leq m \leq Y_{15}$$

is a coefficient 0.95 confidence interval.

(d) (15 points) Finally assume that N is large (e.g. $N > 40$). Use the C.L.T. to find an approximate coefficient γ confidence interval for m.

Solution: For N large a Binomial r.v. with parameters N and 0.5 is close to a Normal r.v. with mean $N/2$ and variance $N/4$ so that we can use the approximation

$$\text{Bin}(i, N, 0.5) \simeq \Phi\left(\frac{2i + 1 - N}{\sqrt{N}}\right)$$

where we have added the "correction to continuity". Thus we need

$$\Phi\left(\frac{2i - 1 - N}{\sqrt{N}}\right) \leq \frac{1 - \gamma}{2}$$

or

$$i \leq \frac{N+1}{2} - \frac{\sqrt{N}}{2}\Phi^{-1}\left(\frac{1-\gamma}{2}\right)$$

Calling $i(N, \gamma)$ the largest integer for which the above inequality holds we have that

$$Y_{i(N,\gamma)} \leq m \leq Y_{N-i(N,\gamma)+1}$$

is an approximate coefficient γ confidence interval.

(e) (20 points (bonus)) Let $\hat{m}(\mathbf{X})$ be the median of the sample defined as

$$\hat{m}(\mathbf{X}) = \begin{cases} Y_{\frac{N+1}{2}} & N \text{ odd} \\ \frac{1}{2}\left(Y_{\frac{N}{2}} + Y_{\frac{N}{2}+1}\right) & N \text{ even.} \end{cases}$$

Show that $\hat{m}(\mathbf{X})$ is a consistent estimator for m. Is it unbiased?

Solution: We will assume that m is unique, that is $F(y)$ is strictly increasing for y near m.

Let $y < m$ so that $F(y) < F(m) = 0.5$. For N odd we have

$$\mathbb{P}(m(\mathbf{X}) < y) = \mathbb{P}(Y_{\frac{N+1}{2}} < y) = \mathbb{P}\left(Q \geq \frac{N+1}{2}\right) = \mathbb{P}\left(Q \geq \frac{N}{2}\right)$$

where Q is a Binomial r.v. with parameters N and $F(y)$, while for N even

$$\mathbb{P}(m(\mathbf{X}) < y) \leq \mathbb{P}(Y_{\frac{N}{2}} < y) = \mathbb{P}\left(Q \geq \frac{N}{2}\right).$$

Observing that

$$\begin{aligned} \mathbb{P}\left(Q \geq \frac{N}{2}\right) =& \mathbb{P}\left(Q - F(y)N \geq \frac{N}{2} - F(y)N\right) \leq \\ & \mathbb{P}\left(|Q - F(y)N| \geq \frac{N}{2} - F(y)N\right) = \\ & \mathbb{P}\left(\left|\frac{Q}{N} - F(y)\right| \geq (1 - 2F(y))\frac{N}{2}\right) \leq \frac{4F(y)(1-F(y))}{((1-2F(y))N)^2}. \end{aligned}$$

so that $\lim_{N\to\infty} \mathbb{P}(m(\mathbf{X}) < y) = 0$.

Similarly since for $Y > m$ we get, for N odd

$$\mathbb{P}(m(\mathbf{X}) > y) = \mathbb{P}\left(Q \leq \frac{N-1}{2}\right) = \mathbb{P}\left(Q \leq \frac{N}{2}\right)$$

while for N even

$$\mathbb{P}(m(\mathbf{X}) > y) \leq \mathbb{P}(Y_{\frac{N}{2}+1} > y) = \mathbb{P}\left(Q \leq \frac{N}{2}\right)$$

and again

$$\mathbb{P}\left(Q \leq \frac{N}{2}\right) \leq \mathbb{P}\left(\left|\frac{Q}{N} - F(y)\right| \geq (2F(y) - 1)\frac{N}{2}\right) \leq \frac{4F(y)(1-F(y))}{((2F(y)-1)N)^2}.$$

Thus for every $y_1 < m < y_2$ we have

$$\lim_{N\to\infty} \mathbb{P}(y_1 < m(\mathbf{X}) < y_2) = 0$$

and thus $m(\mathbf{X})$ is a consistent estimator.

The estimator is clearly not unbiased. If $N = 1$ we have $m(\mathbf{X}) = X_1$ and thus $\mathbb{E}(m(\mathbf{X})) = \mathbb{E}(X_1)$. If X_1 is an exponential r.v. with parameter 1 we have $\mathbb{E}(X_1) = 1$ while $m = \ln(2)$.

Question 2 . . . . *35 point*

Difference between two Normal populations: Let X_i, $i = 1, \ldots, N$ be a random sample from a Normal population with mean μ_X and variance σ_X^2 and Y_i, $i = 1, \ldots, M$ be a random sample from a Normal population with mean μ_Y and variance σ_Y^2. The two samples are independent.

(a) (10 points) Assume that σ_X^2 and σ_Y^2 are known. Find a coefficient γ confidence interval for $\mu_X - \mu_Y$. (**Hint**: $\overline{X}_N$ and $\overline{Y}_M$ are independent Normal r.v.)

Solution: We know that $\overline{X}_N$ is normal with mean μ_X and variance σ_X^2/N while $\overline{Y}_M$ is normal with mean μ_Y and variance σ_Y^2/M. Since they are independent we have that $\overline{X}_N - \overline{Y}_M$ is normal with mean $\mu_X - \mu_Y$ and variance $\sigma_X^2/N + \sigma_Y^2/M$. Thus

$$\frac{\sqrt{N+M}\left(\overline{X}_N - \overline{Y}_M - \mu_X - \mu_Y\right)}{\bar{\sigma}}$$

where

$$\overline{\sigma}^2 = (M+N)(\sigma_X^2/N + \sigma_Y^2/M)$$

is a Normal Standard r.v. and we get that

$$\overline{X}_N - \overline{Y}_M - \frac{\overline{\sigma} z_{\frac{1-\gamma}{2}}}{\sqrt{N+M}} \leq \mu_X - \mu_Y \leq \overline{X}_N - \overline{Y}_M + \frac{\overline{\sigma} z_{\frac{1-\gamma}{2}}}{\sqrt{N+M}}$$

is a coefficient γ confidence interval.

(b) (10 points) Assume now that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ with σ^2 unknown. Assume also that μ_X and μ_Y are unknown. Find a coefficient γ confidence upper limit for σ^2. (**Hint**: use $\sum_{i=1}^{N}(X_i - \overline{X}_N)^2$ and $\sum_{i=1}^{M}(Y_i - \overline{Y}_M)^2$ and the fact that the sum of χ^2 r.v. is a χ^2 r.v.)

Solution: We know that

$$\frac{1}{\sigma^2}\sum_{i=1}^{N}(X_i - \overline{X}_N)^2$$

has a χ^2 distribution with $N-1$ d.o.f. while

$$\frac{1}{\sigma^2}\sum_{i=1}^{M}(Y_i - \overline{Y}_M)^2$$

has a χ^2 distribution with $M-1$ d.o.f. so that

$$\frac{1}{\sigma^2}\left(\sum_{i=1}^{N}(X_i - \overline{X}_N)^2 + \sum_{i=1}^{M}(Y_i - \overline{Y}_M)^2\right)$$

has a χ^2 distribution with $M+N-2$ d.o.f. and the coefficient γ confidence upper limit is

$$\sigma^2 \leq \frac{1}{\chi^2_{\gamma,N+M-2}} \left(\sum_{i=1}^{N} (X_i - \overline{X}_N)^2 + \sum_{i=1}^{M} (Y_i - \overline{Y}_M)^2 \right) .$$

where, if U is a χ^2 r.v. with N d.o.f., we call

$$\mathbb{P}(U \geq \chi^2_{\gamma,N}) = \gamma \, .$$

We can also write it as

$$\sigma^2 \leq \frac{1}{\chi^{-1}_{N+M-2}(1-\gamma)} \left(\sum_{i=1}^{N} (X_i - \overline{X}_N)^2 + \sum_{i=1}^{M} (Y_i - \overline{Y}_M)^2 \right) .$$

where χ_{N+M-2} is the c.d.f. of U.

(c) (15 points) As before assume $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ with σ^2, μ_X, μ_Y unknown. Find a coefficient γ confidence interval for $\mu_X - \mu_Y$. (**Hint**: combines point a) and b) as described in Section 8.4 of the textbook.)

Solution: Calling

$$Z = \frac{\sqrt{NM}\left(\overline{X}_N - \overline{Y}_M - \mu_X - \mu_Y\right)}{\sigma\sqrt{N+M}}$$

and

$$U = \frac{1}{\sigma^2}\left(\sum_{i=1}^{N}(X_i - \overline{X}_N)^2 + \sum_{i=1}^{M}(Y_i - \overline{Y}_M)^2\right)$$

we have that

$$T = \frac{Z}{\sqrt{\frac{U}{M+N-2}}}$$

has a t-distribution with $N + M - 2$ d.o.f. Thus, Calling

$$\Sigma^2 = \frac{1}{N+M-1}\left(\sum_{i=1}^{N}(X_i - \overline{X}_N)^2 + \sum_{i=1}^{M}(Y_i - \overline{Y}_M)^2\right)$$

the coefficient γ confidence interval is

$$\overline{X}_N - \overline{Y}_M - t_{\alpha,N+M-2}\Sigma\sqrt{\frac{1}{N} + \frac{1}{M}}\mu_X - \mu_Y \leq \overline{X}_N - \overline{Y}_M + t_{\alpha,N+M-2}\Sigma\sqrt{\frac{1}{N} + \frac{1}{M}}$$

where $\alpha = (1-\gamma)/2$ and

$$\mathbb{P}(T \geq t_{\alpha,N+M-2}) = \alpha.$$

Alternatively we can write it as

$$\overline{X}_N - \overline{Y}_M - t_{N+M-2}^{-1}(\alpha)\Sigma\sqrt{\frac{1}{N} + \frac{1}{M}}\mu_X - \mu_Y \leq \overline{X}_N - \overline{Y}_M + t_{N+M-2}^{-1}(\alpha)\Sigma\sqrt{\frac{1}{N} + \frac{1}{M}}$$

where t_{N+M-2} is the c.d.f. of T.

Question 3 . . . . *20 point*

Let X_i, $i = 1, \ldots, N$ be a random sample from a population with a unifrom distribution in $[0, A]$.

(a) (10 points) Show that

$$V(\mathbf{X}, A) = \frac{\max_i(X_i)}{A}$$

is a pivotal quantity.

Solution: Observe that X_i/A is uniform in $[0, 1]$ while

$$\frac{\max_i(X_i)}{A} = \max_i \left(\frac{X_i}{A}\right) .$$

Thus the c.d.f. F_V of $V(\mathbf{X}, A)$ is

$$F_V(y) = y^N$$

and does not depend on A. Finally we clearly have

$$A = \frac{\max_i(X_i)}{V(\mathbf{X}, A)} .$$

(b) (10 points) Use the pivotal quantity $V(\mathbf{X}, A)$ to create a coefficient γ confidence interval for A.

Solution: Calling $\alpha = (1 - \gamma)/2$, theorem 8.5.3 in the textbook tell us that

$$\frac{\max_i(X_i)}{(1-\alpha)^{\frac{1}{N}}} \leq A \leq \frac{\max_i(X_i)}{\alpha^{\frac{1}{N}}}$$

is a coefficient γ confidence interval.