415 Midterm 1

1. Do not open until told to do so. When you start check that you have 3 sheets, 6 pages of problems.
2. Do not write your response on the problem sheet. Write it on the blank scratch paper.
3. Write your name below in print.
4. Staple problem sheet and response together and submit.
5. Make sure you submit in the right pile (MATH STAT).
6. Keep your cheat sheet for use in future exams (not mandatory).
7. Pick up your HW1 to HW4 if you have not done so already. I will discard them otherwise.

MATH or STAT (circle one)

Name: ______________________________

1. (18 pts) Answer T (true) or F (false) to the following statements. No need to show intermediate derivations.

1. X has continuous cdf $F_X(x)$ on $[a, b]$. Then $F_X(X)$ follows $U(0, 1)$. T
2. If $Z \sim N(0, 1)$ and F is continuous cdf strictly increasing on $[a, b]$. Then $F^{-1}(Z)$ is continuous r.v. with cdf F. F
3. $E(X^2) = [E(X)]^2$. F
4. X_1 up to X_n are i.i.d. normal, $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is unbiased for σ^2. T
5. Z_1 up to Z_n are i.i.d. standard normal, $Z_1^2 + \cdots Z_n^2$ is $\chi^2(n)$. T
6. X_1 up to X_n are independent normal, c_1 up to c_n are arbitrary constants, $\sum_{i=1}^{n} c_i X_i$ is not necessarily normal. F
7. t_r for $r < \infty$ has a fatter tail than a standard normal. T
8. The Central Limit Theorem can only be used on a sample of i.i.d. distributions that are not normal. F
9. If you had a higher batting average than me for both the first half of the season and the last half of the season, you will have a higher batting average than me for the full season. F
10. The first order statistic and the nth order statistic of a size n sample are respectively the min and max of the sample. T
11. Maximum Likelihood estimators are always unbiased. F
12. $E\left[\frac{1}{\bar{X}}\right] = \frac{1}{E[\bar{X}]}$. F
13. To get the distribution of $\frac{1}{\bar{X}}$, you would first need to find the distribution of $\bar{X}$ and then perform a change of variables. T
14. The Maximum likelihood estimator of θ for a sample X_1 up to X_n i.i.d. from $U(0, \theta)$ is the maximum of the sample. T
15. If you can write a pdf of a random variable in an exponential form $\exp[a(x)p(\theta) + b(x) + q(\theta)]$, $a(x)$ is a sufficient statistic for θ. T
16. $E[E[X|Y]] = E[Y]$. F
17. Two different random variables can have the same MGF. F
18. Z does not involve the parameters θ, it is therefore ancillary. T

2. (32 pts) For each of the following problems, answer True or False to each option. No need to show intermediate derivations.

1. For i.i.d. normal X_1 up to X_n,
   (a) about 90% of X_i's lie in $(\bar{X} - 2S, \bar{X} + 2S)$. F
   (b) about 95% of X_i's lie in $(\bar{X} - 2S, \bar{X} + 2S)$. T
   (c) about 99.7% of X_i's lie in $(\bar{X} - S, \bar{X} + S)$. F
   (d) 100% of X_i's lie in $(\bar{X} - S, \bar{X} + S)$. F

2. The inter-quartile range spans from,
   (a) the first quartile to the second quartile. F
   (b) the second quartile to the thrid quartile. F
   (c) the first qurtile to the thrid quartile. T
   (d) the min to the max. F

3. In a Q-Q plot, for a sample of size n, X_1 up to X_n, the sample quantile $X_{(r)}$ is plotted against the theoretical quantile $q_{\frac{r}{n+1}}$,
   (a) if the sample follows the theoretical distribution, the points will lie on the straight line with slope 1 and intercept 0. T
   (b) you can also plot the theoretical quantiles against the sample quantiles. T
   (c) you will need to calculate q_0. F
   (d) in $q_{\frac{r}{n+1}}$, we take division by $n + 1$ instead of n because if we take division by n, we get q_1 for $r = n$ which might be ∞. T

4. We want to find the θ that maximizes the function $L(\theta, X)$,
   (a) you would always need to take the derivative of $L(\theta, X)$. F
   (b) you can find θ that maximizes $\log L(\theta, X)$ instead. T
   (c) if $L(\theta, X) = M(\theta, X)N(X) + Q(X)$ you can maximize $M(\theta, X)$ instead. F
   (d) if $L(\theta, X) = M(\theta, X)^{R(X)}$ you can maximize $M(\theta, X)$ instead (note sign of $R(X)$). F

5. $V(\hat{\beta}|X) = \frac{1}{n}\frac{1}{\left[\bar{X^2} - (\bar{X})^2\right]}\sigma^2$ for least squares regression with model $Y = \alpha + \beta X + \epsilon$, $V(\epsilon) = \sigma^2$,
   (a) the less samples you have, the better the estimate $\hat{\beta}$. F
   (b) the less variable X is, the better the estimate $\hat{\beta}$. F
   (c) given $(\bar{X})^2$ fixed, the smaller the $\bar{X^2}$, the better the estimate $\hat{\beta}$. F
   (d) given $\bar{X^2}$ fixed, the smaller the $(\bar{X})^2$, the better the estimate $\hat{\beta}$. T

6. $X_1, \cdots, X_n$ random sample, $f(x_1, \cdots, x_n, \theta) = f_1(y, \theta) f_2(x_1, \cdots, x_n)$, $Y = u(X_1, \cdots, X_n)$, big letters are the random samples, small letters are the fixed realizations,

   (a) Y is a sufficient statistic for θ. T

   (b) Y^2 is sufficient statistic for θ. F

   (c) Y^3 is sufficient statistic for θ. T

   (d) e^Y is sufficient statistic for θ. T

7. **Rao-Blackwell theorem** Y is sufficient statistic for θ, Z is unbiased estimator of θ,

   (a) $E[Z|Y]$ involves θ. F

   (b) $E[Z|Y]$ is biased. F

   (c) $E[Z|Y]$ has larger variance than Z. F

   (d) $E[Z] = \theta$. T

8. **Basu's theorem** Y is sufficient statistic for θ, Z is ancillary statistic,

   (a) if density of Y is exponential form, Z and Y are independent. T

   (b) if Y is complete, Z and Y are independent. T

   (c) Laplace transform maps f to $F(s) = \int_{\mathbb{R}} e^{-st} f(t) dt$, if $g = 2f$, $G(s) = F(s)$. F

   (d) A statistic T is complete if for any g, we have that if $E_\theta[g(T)] = 0$ for all θ, then $P[g(T) = 1] = 1$. F

3. (5 pts) X_1, X_2, X_3 are i.i.d. from $U(0,1)$. Find the pdf of the median. Don't just write the answer. Derive it. $\left(\binom{3}{2}x^2(1-x)+x^3\right)'$.

4. (4 pts) For above you should get $6x(1-x)$ on $(0,1)$. Find $E[X_{(2)}]$. $\int_0^1 6x^2(1-x)$.

5. (6 pts) We have n i.i.d. samples from a distribution with density $2\theta x e^{-\theta x^2}$ on x greater than 0. Find maximum likelihood estimator of θ. $L = 2^n\theta^n \Pi_{i=1}^n X_i e^{-\theta\sum_{i=1}^n X_i^2}$ so $l = n\log 2 + n\log\theta + \sum_{i=1}^n \log X_i - \theta\sum_{i=1}^n X_i^2$. So $\frac{dl}{d\theta} = \frac{n}{\theta} - \sum_{i=1}^n X_i^2 = 0$. So MLE is $\frac{n}{\sum_{i=1}^n X_i^2}$.

6. (5 pts) We have data $Y_1 = 1$, $Y_2 = -1$ and $X_1 = 2$ $X_2 = 4$. We think of the following model $Y_i = \alpha + \beta X_i + \epsilon_i$. ϵ_i's are the error term. Find $\hat\alpha$ and $\hat\beta$ that minimizes

$$(Y_1 - \hat\alpha + \hat\beta X_1)^2 + (Y_2 - \hat\alpha + \hat\beta X_2)^2.$$

$\hat\alpha = 3$. $\hat\beta = -1$.

7. (4 pts) Standardize the X_i's by $\tilde X_i = \frac{X_i - 3}{2}$. What is the estimate of the intercept and slope for regression of Y_i's on the $\tilde X_i$'s? $\hat\alpha = 0$. $\hat\beta = -2$.

8. (3 pts) In problem 6, will the estimates change if you were minimizing

$$|Y_1 - \hat\alpha + \hat\beta X_1| + |Y_2 - \hat\alpha + \hat\beta X_2|.$$

instead? No.

9. (5 pts) In least squares linear regression, $e_i = Y_i - \hat\alpha - \hat\beta X_i$, $\hat\beta = \frac{\bar{XY} - \bar X\bar Y}{\bar{X^2} - (\bar X)^2}$, $\hat\alpha = \bar Y - \bar X\hat\beta$. Show that

$\sum_{i=1}^{n} e_i X_i = 0$.

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i)X_i &= \sum_{i=1}^{n}(Y_iX_i - \hat{\alpha}X_i - \hat{\beta}X_i^2) \\
&= n\bar{XY} - n\hat{\alpha}\bar{X} - n\hat{\beta}\bar{X^2} = n\bar{XY} - n(\bar{Y} - \bar{X}\hat{\beta})\bar{X} - n\hat{\beta}\bar{X^2} \\
&= n\bar{XY} - n\bar{Y}\bar{X} + n(\bar{X})^2\hat{\beta} - n\hat{\beta}\bar{X^2} \\
&= n\bar{XY} - n\bar{Y}\bar{X} - \hat{\beta}\left[n\bar{X^2} - n(\bar{X})^2\right] \\
&= n\bar{XY} - n\bar{Y}\bar{X} - \frac{n\bar{XY} - n\bar{X}\bar{Y}}{n\bar{X^2} - n(\bar{X})^2}\left[n\bar{X^2} - n(\bar{X})^2\right] = 0.
\end{aligned}$$

There are other easier ways.

10. (5 pts) We have a i.i.d. sample of size n from a Beta distribution with density

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

on $(0,1)$. Say α is known. Find the sufficient statistic for β. $\Pi_{i=1}^{n}(1-X_i)$.

11. (5 pts) Show that the probability mass function of the Binomial distribution

$$\binom{n}{x}p^x(1-p)^{n-x} \ (x = 0, \cdots, n)$$

is of exponential form.

$$\begin{aligned}
\exp\left(\log\binom{n}{x} + x\log p + (n-x)\log(1-p)\right) &= \exp\left(\log\binom{n}{x} + x\log p + n\log(1-p) - x\log(1-p)\right) \\
&= \exp\left(\log\binom{n}{x} + x\log\frac{p}{1-p} + n\log(1-p)\right).
\end{aligned}$$

12. (4 pts) Show that $E[YX|X] = XE[Y|X]$. Set $T = XY$, we have

$$\int_{-\infty}^{\infty} t f_{T|X=x}(t|x)dt.$$

perform change of variables from T to Y.

$$\int_{-\infty}^{\infty} xyf_{XY|X=x}(xy|x)xdy = \int_{-\infty}^{\infty} xy\frac{1}{x}f_{Y|X=x}(y|x)xdy$$

Note that $f_{XY|X=x}(xy|x) = \frac{1}{x}f_{Y|X=x}(y|x)$. This is because for X with density $f(x)$, cX has density $\frac{1}{c}f(cx/c)$. This arrive from the change of variables from X to cX.

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} f(cx/c)\frac{1}{c}d(cX).$$

13. (4 pts) From the equality

$$V[Z] = V[E[Z|Y]] + E[V[Z|Y]]$$

argue that

$$V[E[Z|Y]] \leq V[Z].$$

$V[Z|Y]$ is always greater than or equal to 0. It's a random variable but the expected value is also greater than or equal to 0.