Math 583 Final, Fall 2020 Name_____

1) Consider the following ordered data set.

1.0 1.8 2.0 2.2 2.3 2.3 2.4 2.8 3.4 5.6

a) Find the sample median MED(n).

b) Find the sample median absolute deviation MAD(n).

Parts c)-f) refer to the CI based on MED(n).

c) Find L_n.

d) Find U_n.

e) Find the degrees of freedom p.

f) Find SE(MED(n)).

```
Coefficient Estimates Response   Y =1 if ape, Y =0 if human
Label      Estimate        Std. Error     Est/SE    p-value
Constant     11.5092        5.46270         2.107     0.0351
lower jaw   -0.360127       0.132925       -2.709     0.0067
upper jaw    0.779162       0.382219        2.039     0.0415
face length -0.374648       0.238406       -1.571     0.1161
```

2) Use the above output to predict the probability $\hat{\rho}(\boldsymbol{x})$ (that a skeleton is ape or human) if *lower jaw* $= x_2 = 97$, *upper jaw* $= x_3 = 76$ and *face length* $= x_4 = 115$.

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.901459   0.186877  10.175  0.000
t            0.556003   0.045780  12.145  0.000
t^2         -0.021346   0.002659  -8.029  0.000
```

3) The above Poisson regression output is for $Y =$ *number* of new AIDS cases in Belgium from 1981 to 1993. The variable $x_2 = t =$ *time* was coded as 1 for 1981, 2 for 1982, ..., 13 for 1993. There were $n = 13$ cases. Predict $\hat{Y} = \hat{\mu}(\boldsymbol{x})$ if $t = x_2 = 7.0$ and $x_3 = t^2 = 49.0$.

crancap	hdlen	hdht
1485	175	132
1450	191	117
1460	186	122
1425	191	125
1430	178	120
1290	180	117
90	75	51

4) The above table represents 3 head measurements on 6 people and one ape. Let X_1 = *cranial capacity*, X_2 = *head length* and X_3 = *head height*. Let $\boldsymbol{x} = (X_1, X_2, X_3)^T$. Several multivariate location estimators, including the coordinatewise median and sample mean, are found by applying a univariate location estimator to each random variable and then collecting the results into a vector.

a) Find the coordinatewise median MED($\boldsymbol{x}$).

b) Find the sample mean $\overline{\boldsymbol{x}}$.

5) If Y has a one sided stable distribution, $Y \sim \text{OSS}(\sigma)$, then the population median $\text{MED}(Y) = \sigma/0.4549$. If you have a random sample $Y_1, ..., Y_n$ which are iid $\text{OSS}(\sigma)$, suggest a robust estimator for σ based on the sample median MED(n) and or the sample MAD(n).

6) The table below shows simulation results for bootstrapping OLS (reg), lasso, and ridge regression (RR) with 10-fold CV when $\boldsymbol{\beta} = (1, 1, 0, 0)^T$. The β_i columns give coverage = the proportion of CIs that contained β_i and the average length of the CI. The test is for $H_0 : (\beta_3, \beta_4)^T = \mathbf{0}$ and H_0 is true. The "coverage" is the proportion of times the prediction region method bootstrap test failed to reject H_0. OLS used 1000 runs while 100 runs were used for lasso and ridge regression. Since 100 runs were used, a cov in [0.89, 1] is reasonable for a nominal value of 0.95. If the coverage for both methods ≥ 0.89, the method with the shorter average CI length was more precise. (If one method had coverage ≥ 0.89 and the other had coverage < 0.89, we will say the method with coverage ≥ 0.89 was more precise.) (Lengths for the test column are not comparable unless the statistics have the same asymptotic distribution.)

Table 1: Bootstrapping lasso and RR, $n = 100, \psi = 0, p = 4, B = 250$

		β_1	β_2	β_3	β_4	test
reg	cov	0.945	0.947	0.941	0.941	0.937
	len	0.397	0.399	0.400	0.398	2.451
RR	cov	0.95	0.89	0.95	0.95	0.94
	len	0.401	0.366	0.377	0.382	2.451
reg	cov	0.928	0.948	0.953	0.952	0.943
	len	0.661	0.673	0.675	0.676	2.490
lasso	cov	0.97	0.90	0.99	0.98	0.97
	len	0.684	0.741	0.612	0.610	2.650

a) For β_3 and β_4 which method, ridge regression or the OLS full model, was more precise?

b) For β_3 and β_4 which method, lasso or the OLS full model, was more precise?

7) Suppose you are estimating the mean μ of losses with $T = \overline{X}$.
actual losses 1, 2, 5, 10, 50: $\overline{X} = 13.6$,
a) Compute $T_1^*, ..., T_4^*$, where T_i^* is the sample mean of the ith bootstrap sample.
bootstrap samples:

2, 10, 1, 2, 2:

50, 10, 50, 2, 2:

10, 50, 2, 1, 1:

5, 2, 5, 1, 50:

b) Now compute the bagging estimator which is the sample mean of the T_i^*: the bagging estimator $\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^*$ where $B = 4$ is the number of bootstrap samples.

8) Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors where

$$\boldsymbol{x}_i \sim (1-\gamma)N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \gamma N_p(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$$

with $0 < \gamma < 1$ and $c > 0$. Then $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{x}_i) = [1 + \gamma(c-1)]\boldsymbol{\Sigma}$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{d})$ for appropriate vector $\boldsymbol{d}$.

9) Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left(\begin{pmatrix} 1 \\ 7 \\ 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 0 & 2 & 1 \\ 0 & 1 & 0 & 0 \\ 2 & 0 & 3 & 1 \\ 1 & 0 & 1 & 5 \end{pmatrix} \right).$$

a) Find the distribution of X_3.

b) Find the distribution of $(X_1, X_4)^T$.

c) Which pairs of random variables X_i and X_j are independent?

d) Find the correlation $\rho(X_1, X_3)$.

10) Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. Find the distribution of $\boldsymbol{A}\boldsymbol{Y}$ if $\boldsymbol{A}$ is an $p \times n$ full rank constant matrix.

11) The data below are a sorted residuals from a lasso regression where $n = 1000$ and $p = 17$. Find shorth(997) of the residuals.

number	1	2	3	4	...	997	998	999	1000
residual	-3.28	-3.06	-3.04	-2.96	...	2.66	2.71	2.81	3.62

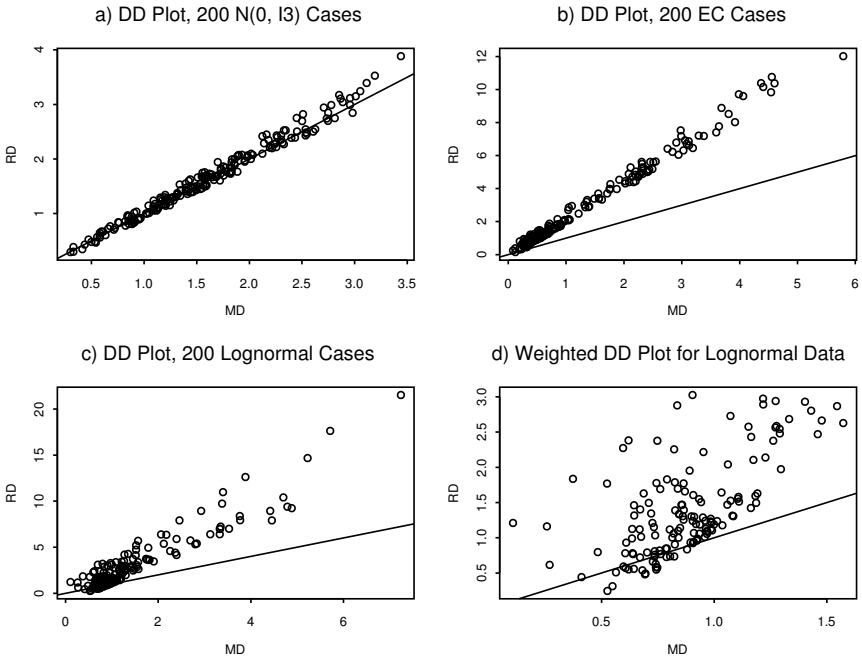

Figure 1: 4 DD Plots

12) Refer to the above DD plots.

a) Why would you conclude the data in DD plot a) come from a multivariate normal distribution?

b) Why would you conclude the data in DD plot b) come from an elliptically contoured distribution that is not a multivariate normal distribution? distribution?