STAT 305 D Exam 1

Show all your work.

1. (20 points) Caustic stress corrosion cracking of iron and steel has been studied because of failures around rivets in steel boilers and failures of steam rotors. A new teflon coating may reduce the corrosion behind this cracking. 10 smooth iron bars and 10 smooth steel bars were taken for the study. 5 iron bars were given the teflon coating, and 5 iron bars were left bare. Similarly, 5 steel bars were given the teflon coating and the remaining 5 were left bare. Constant load stress corrosion tests (with constant load and constant stress) were applied for the same length of time to each bar. The length of the longest crack in μm was measured for each.

   a. (3 points) Identify the sample (or samples).

      - Sample 1: the 10 steel bars.
      - Sample 2: the 10 iron bars.

   b. (3 points) Identify the population (or populations).

      - Population 1: All the steel bars that could have been selected for the study.
      - Population 2: All the iron bars that could have been selected for the study.

   c. (3 points) Identify and classify all the variables.

      - Treatment variable: teflon or no teflon
      - Blocking variable: steel or iron
      - Response variable: length of longest crack

   d. (3 points) Is this study an experimental study or an observational study?

      The study is an experiment because the experimenters actively control the treatment variable of interest (teflon or no teflon) and hold the experimental conditions were either constant or randomized across levels of teflon.

   e. (4 points) Suppose the teflon-coated bars corrode and crack less for both steel and iron bars. Can we say that the teflon *causes* this reduction in corrosion and cracking? Why or why not?

      Yes: the experimenters set the levels of teflon themselves, and all the experimental conditions were either constant or randomized across levels of teflon.

   f. (4 points) Suppose the steel bars corrode and crack less than the iron bars in the study. Can we say that the choice of steel over iron causes this difference in corrosion and cracking? Why or why not?

No: since the experimenters could not randomly assign bars to different materials (each bar was either already steel or already iron), the experimental conditions could have been correlated with bar material. For example, steel bars could have been brand new and the iron bars could have been old and rusty.

2. (20 points)

   a. (10 points) Using the table of random digits below, select a simple random sample of 10 steel bars from a shipment of 100 steel bars for the study in question 1. Also, select a simple random sample of 10 iron bars from a shipment of 50 iron bars for the study, continuing in the table of random digits from where you left off from the steel bars. Carefully describe how you did this.

Random Digits

12159	66144	05091	13446	45653	13684	66024	91410	51351	22772
30156	90519	95785	47544	66735	35754	11088	67310	19720	08379
59069	01722	53338	41942	65118	71236	01932	70343	25812	62275
54107	58081	82470	59407	13475	95872	16268	78436	39251	64247
99681	81295	06315	28212	45029	57701	96327	85436	33614	29070

   - Steel bars: first, I assign a 2-digit index to each steel bar in the shipment: 00, 01, ..., 99. Next, I move along the top row of the table of random digits from left to right, selecting bars 12, 15, 96, 61, 44, 05, 09, 11, 34 and 46 for the study.
   - Iron bars: first, I assign a 2-digit index to each iron bar in the shipment: 00, 01, ..., 49. Next, I continue where I left off in the random number table, moving left to right and selecting 45, 31, 36, 02, 49, 14, 10, 35, 12, and 27 for the study.

   b. (10 points) Using a different table of random digits (below), randomize the 10 steel bars to receive teflon or remain bare. Then, randomize the 10 iron bars to receive teflon or remain bare, continuing in the table of random digits from where you left off from the steel bars. Carefully describe how you did this.

   - First, I assign the 10 steel bars in the sample an index from 0 to 9. Then, I move along the top row of the table table from left to right, selecting bars 2, 7, 5, 3, and 8 to receive teflon. The other bars will remain bare.
   - First, I assign the 10 iron bars in the sample an index from 0 to 9. Then, I move along the top row of the table table from left to right, continuing where I left off from the steel bars. I select bars 7, 5, 3, 6, and 9 to receive teflon. The other bars will remain bare.

27252	37875	53679	01889	35714	63534	63791	76342	47717	73684
93259	74585	11863	78985	03881	46567	93696	93521	54970	37601
84068	43759	75814	32261	12728	09636	22336	75629	01017	45503
68582	97054	28251	63787	57285	18854	35006	16343	51867	67979
60646	11298	19680	10087	66391	70853	24423	73007	74958	29020

3. (20 points) Revisit the New York Rivers data from class.

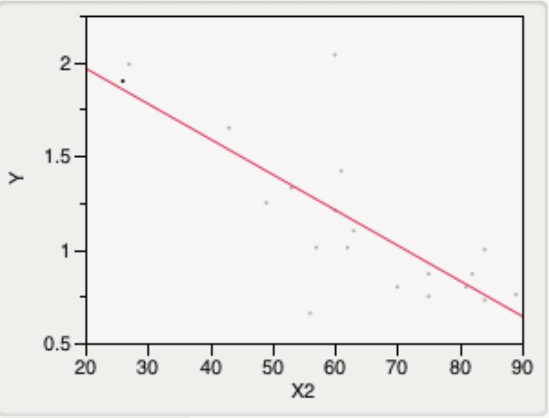
Name	X1	X2	X3	X4	Y
Olean	26	63	1.20	0.29	1.10
Cassadaga	29	57	0.70	0.09	1.01
Oatka	54	26	1.80	0.58	1.90
Neversink	2	84	1.90	1.98	1.00
Hackensack	3	27	29.40	3.11	1.99
Wappinger	19	61	3.40	0.56	1.42
Fishkill	16	60	5.60	1.11	2.04
Honeoye	40	43	1.30	0.24	1.65
Susquehanna	28	62	1.10	0.15	1.01
Chenango	26	60	0.90	0.23	1.21
Tioughnioga	26	53	0.90	0.18	1.33
West_Canada	15	75	0.70	0.16	0.75
East_Canada	6	84	0.50	0.12	0.73
Saranac	3	81	0.80	0.35	0.80
Ausable	2	89	0.70	0.35	0.76
Black	6	82	0.50	0.15	0.87
Schoharie	22	70	0.90	0.22	0.80
Raquette	4	75	0.40	0.18	0.87
Oswegatchie	21	56	0.50	0.13	0.66
Cohocton	40	49	1.10	0.13	1.25

Remember:

- Y is the mean nitrogen content (mg/liter).
- X1 is the percent agricultural land
- X2 is the percent forested land
- X3 is the percent residential land
- X4 is the percent commercial/industrial land

Below, I fit a regression line of Y on X2.

Bivariate Fit of Y By X2



Linear Fit

Linear Fit

Y = 2.3471305 - 0.0189281*X2

Summary of Fit

RSquare	0.598201
RSquare Adj	0.575879
Root Mean Square Error	0.284365
Mean of Response	1.1575
Observations (or Sum Wgts)	20

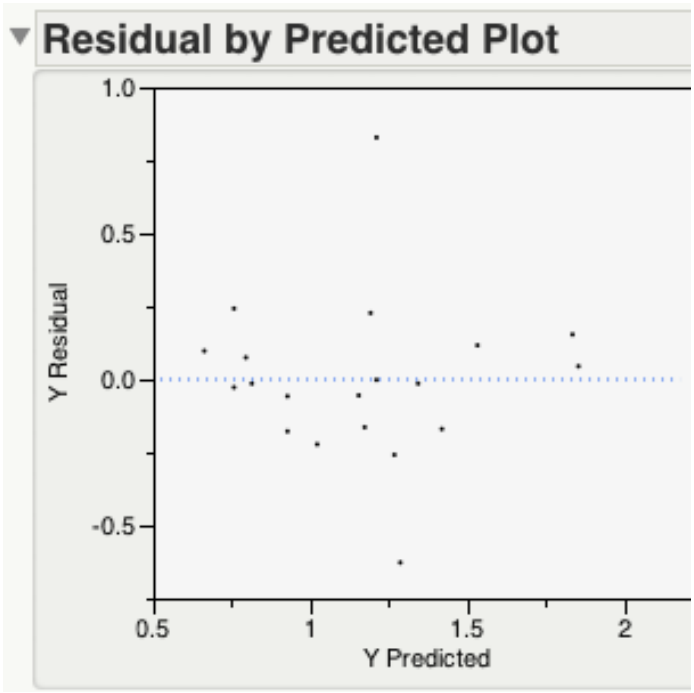
Lack Of Fit

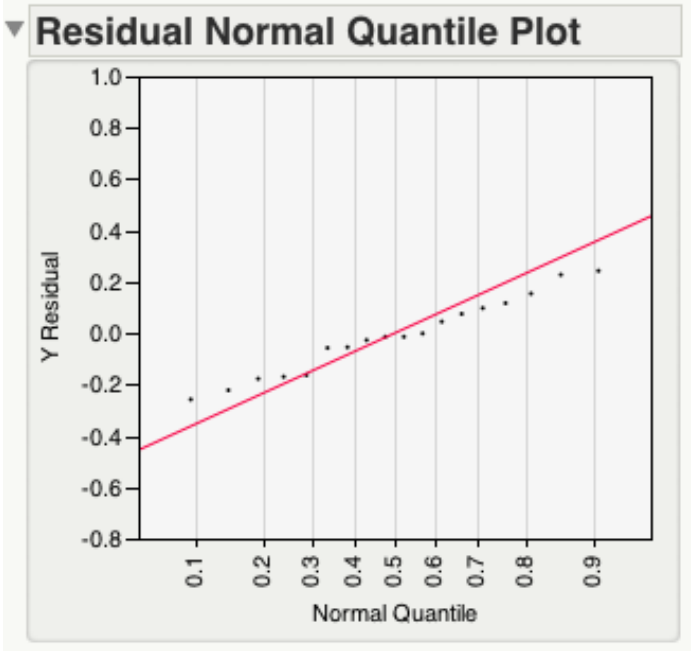
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	2.1670297	2.16703	26.7986
Error	18	1.4555453	0.08086	Prob > F
C. Total	19	3.6225750		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.3471305	0.238438	9.84	<.0001*
X2	-0.018928	0.003656	-5.18	<.0001*





a. (4 points) Identify and interpret the slope.
The slope is -0.0189 gm/L/% forested land. On average, the mean nitrogen content decreases by -0.0189 gm/L for every 1% increase in

surrounding forested land.

b. (4 points) Identify and interpret the intercept.
The intercept is 2.347 mg/L. The model predicts that on average, a river with no surrounding forested land should have a mean nitrogen content of 2.347 mg/L.

c. (4 points) The intercept predicts the nitrogen content for a river with 0% surrounding forested land. Why might this prediction be a bad idea in practice?
We have no data from rivers with 0% forested land. In fact, the lowest percentage of forested land we have for a data point is 20%. Attempting to predict at X2 = 0% would be extrapolating far beyond the range of the data, a dangerous practice.

d. (4 points) Based on the residual plot, comment on the validity of the model.
. Since there is no apparent pattern in the residuals, the model appears valid.

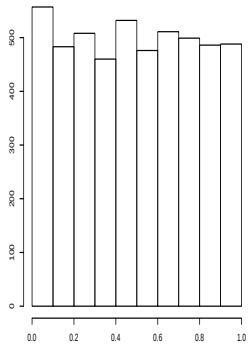
e. (4 points) Based on the normal quantile (normal QQ) plot, do the residuals look bell-shaped (normally-distributed)?
Since the points in the normal QQ plot appear as a straight line, the residuals look normally distributed.

4. (20 points)

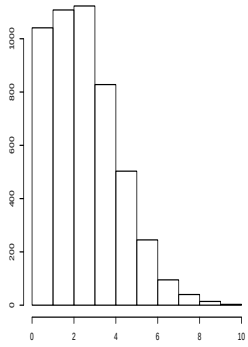
a. (10 points) What is the difference between a histogram and a bar plot?
A histogram displays continuous numerical data by dividing the data into equally-sized intervals before plotting. A bar plot displays discrete or categorical data.

b. (10 points) Identify the following distributional shapes.

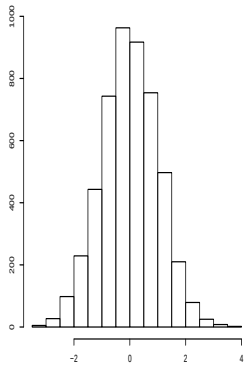
i. (2.5 points) uniform



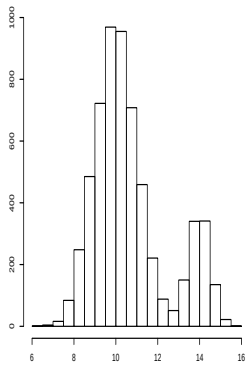
ii. (2.5 points) skewed right



iii. (2.5 points) bell-shaped



iv. (2.5 points) bimodal, asymmetric



5. (20 points) The article "Effects of Aggregates and Microfillers on the Flexural Properties of Concrete (Magazine of Concrete Research, 1997: 8198) reported on a study of strength properties of high-performance concrete obtained by using superplasticizers and certain binders. The compressive strength of such concrete had previously been investigated, but not much was known about flexural strength (a measure of ability to resist failure in bending). The accompanying data on flexural strength (in MegaPascal, MPa, where 1 Pa (Pascal) = 1.45 $\times 10^4$ psi) is part of the data that appeared in the article cited:

7.0 7.4 7.7 7.8 7.9 8.1 8.7 9.0 9.7 11.3 11.8

a. (10 points) Find $Q(0.25)$ and $Q(0.75)$ of the data.

Data	7.000	7.400	7.700	7.800	7.900	8.100	8.700	9.000	9.700	11.300	11.800
i	1.000	2.000	3.000	4.000	5.000	6.000	7.000	8.000	9.000	10.000	11.000
$\frac{i-.5}{11}$	0.045	0.136	0.227	0.318	0.409	0.500	0.591	0.682	0.773	0.864	0.955

For $Q(0.25)$:

$$\begin{aligned} i' &= np + 0.5 \\ &= 11 \cdot 0.25 + 0.5 \\ &= 3.25 \end{aligned}$$

Hence:

$$\begin{aligned} Q(0.25) &= (\lceil i' \rceil - i')x_{\lfloor i' \rfloor} + (i' - \lfloor i' \rfloor)x_{\lceil i' \rceil} \\ &= (\lceil 3.25 \rceil - 3.25)x_{\lfloor 3.25 \rfloor} + (3.25 - \lfloor 3.25 \rfloor)x_{\lceil 3.25 \rceil} \\ &= (4 - 3.25)x_3 + (3.25 - 3)x_4 \\ &= 0.75 \cdot 7.7 + 0.25 \cdot 7.8 \\ &= 7.725 \end{aligned}$$

For $Q(0.75)$:

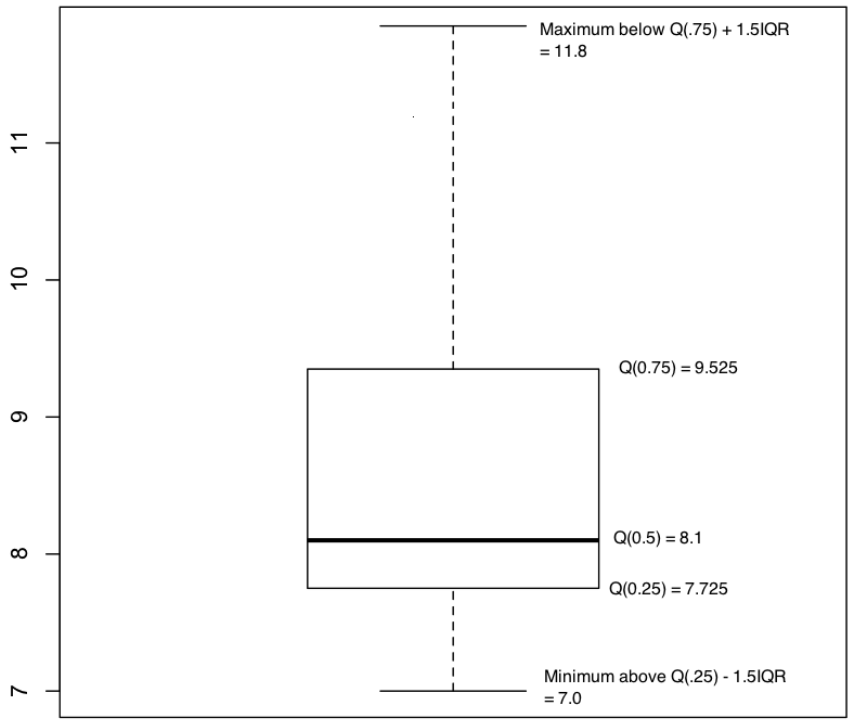
$$\begin{aligned} i' &= np + 0.5 \\ &= 11 \cdot 0.75 + 0.5 \\ &= 8.75 \end{aligned}$$

Hence:

$$\begin{aligned} Q(0.75) &= (\lceil i' \rceil - i')x_{\lfloor i' \rfloor} + (i' - \lfloor i' \rfloor)x_{\lceil i' \rceil} \\ &= (\lceil 8.75 \rceil - 8.75)x_{\lfloor 8.75 \rfloor} + (8.75 - \lfloor 8.75 \rfloor)x_{\lceil 8.75 \rceil} \\ &= (9 - 8.75)x_8 + (8.75 - 8)x_9 \\ &= 0.25 \cdot 9 + 0.75 \cdot 9.7 \\ &= 9.525 \end{aligned}$$

b. (10 points) Make a boxplot of the data. Is the distribution symmetric? Are there any outliers?

- The median, $Q(0.5)$, is the middle data point, $x_6 = 8.1$.
- 1.5 IQR $= 1.5(9.525 - 7.725) = 2.7$
- $Q(0.25) - 1.5IQR = 7.725 - 2.7 = 5.025$
- $Q(0.75) + 1.5IQR = 9.525 + 2.7 = 12.225$
- There are no points below $Q(0.25) - 1.5IQR$ or above $Q(0.75) + 1.5IQR$, so there are no outliers. The boxplot is shown below. The distribution looks skewed up (skewed right), not symmetric.


Maximum below Q(.75) + 1.5IQR
= 11.8
Q(0.75) = 9.525
Q(0.5) = 8.1
Q(0.25) = 7.725
Minimum above Q(.25) - 1.5IQR
= 7.0
11
10
9
8
7