ST 305: Final Exam

By handing in this completed exam, I state that I have neither given nor received assistance from another person during the exam period. I have not copied from another person's paper. I have used no resources other than the exam itself and the basic mathematical functions of a calculator (ie, no notes, electronic communication, notes stored in calculator memory, etc.). I have not used my calculator to compute values from statistical functions such as the standard normal or t. I understand that the penalty if I am found guilty of any such cheating will include failure of the course and a report to the NCSU Office of Student Conduct. **I understand that I must show all work/calculations, even if they seem trivial, to get credit for my answers. Answers without proper justification/defense will get no or reduced credit.**

Name: KEY

ID#: ____________________

$$\bar{x} = \frac{1}{n}\sum x_i$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$Z = \frac{X - \mu}{\sigma}$$

$$r = \frac{\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1}$$

$$b_1 = r\frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$\text{residual} = y - \hat{y}$$

$$P(A \text{ or } B) = P(A) + P(B)$$

$$P(A^C) = 1 - P(A)$$

$$P(A \text{ and } B) = P(A) \times P(B)$$

$$\mu_X = \sum x_i p_i$$

$$\mu_{a+bX} = a + b\mu_X$$

$$\mu_{X+Y} = \mu_X + \mu_Y$$

$$\sigma_X^2 = \sum (x_i - \mu_X)^2 p_i$$

$$\sigma_{a+bX}^2 = b^2\sigma_X^2$$

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$$

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2$$

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y$$

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$$P(A \text{ and } B) = P(A)P(B \mid A)$$

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

$$\mu_X = np$$

$$\sigma_X = \sqrt{np(1-p)}$$

$$\hat{p} = X / n$$

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$$P(X = k) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \sigma/\sqrt{n}$$

$$m = z^*\sigma/\sqrt{n}$$

$$\bar{x} \pm m$$

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

$$n = \left(\frac{z^*\sigma}{m}\right)^2$$

Simple Linear Regression

$$b_1 = r\frac{s_y}{s_x}; \quad b_0 = \bar{y} - b_1\bar{x}$$

$$e_i = y_i - \hat{y}_i; \quad s^2 = \frac{\sum e_i^2}{n-2}$$

$$b_j \pm t^* SE_{b_j}; \quad t = \frac{b_j}{SE_{b_j}} \quad \text{df} = n-2$$

$$\hat{\mu}_y \pm t^* SE_{\hat{\mu}}; \quad \hat{y} \pm t^* SE_{\hat{y}}$$

$$\text{SST} = \sum (y_i - \bar{y})^2$$

$$\text{SSM} = \sum (\hat{y}_i - \bar{y})^2$$

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2$$

$$SE_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

$$SE_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

$$SE_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$SE_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Multiple regression changes:

$$s^2 = \frac{\sum e_i^2}{n-p-1}$$

$$\text{df} = n-p-1$$

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}, \ \text{df} = n-1$$

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \ \text{df} = \min(n_1, n_2) - 1$$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \ \text{df} = n-1$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \ \text{df} = \min(n_1, n_2) - 1$$

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \ \text{df} = n_1 + n_2 - 2$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \ \text{df} = n_1 + n_2 - 2$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\psi = \sum a_i \mu_i, \quad c = \sum a_i \bar{x}_i$$

$$SE_c = s_p \sqrt{\sum \frac{a_i^2}{n_i}}$$

Definitions. (1 point each)

Clearly define each of the following terms.

1. Distribution: The possible values of a variable, and the frequencies of each value

2. Correlation: A numerical measure of the strength & direction of a linear relationship between 2 variables

3. Parameter: A numerical property of a population

4. Unbiased estimate: An estimate which has an expected valu equal to the parameter being estimated

5. Margin of Error: A value reflecting the accuracy of an estimate at a chosen level of confidence

6. Double blind experiment: An experiment where neither subjects nor experimenters know which treatments individual subjects received

7. Simple Random Sample: Sample selected in such a way that every combination of n individuals has an equal chance of selection

8. Power (of a test): The probability of correctly rejecting H_0

9. Central Limit Theorem: For a SRS of size n from any pop'n with mean μ & std dev σ, $\bar{x} \dot{\sim} N(\mu, \sigma/\sqrt{n})$ for large n

10. Type I Error:

The probability of rejecting H_0 when it is actually correct.

Pick the Procedure. (2 points each)

Select the procedure that would be the best approach to answer each of the following questions. You only need to give the letter of the answer (**please write it clearly to the left of each question, or lose 1 point**). CI: confidence interval, HT: hypothesis test

a. CI for a single mean
b. HT for a single mean
c. CI for a difference in two means
d. HT for a difference in two means
e. Matched Pairs CI
f. Matched Pairs HT
g. Simple linear regression
h. Multiple linear regression
i. 1-Way ANOVA -~~CI~~
j. 2-Way ANOVA- ~~HT~~

I 11. Are there differences in the number of hours per week freshmen, sophomores, juniors, and seniors spend working at part-time jobs?

G 12. Do Congressional committees with more members propose more laws?

H 13. How well does a collection of six measures of health predict a person's lifespan?

F 14. Does the number of drug side effects decrease after a patient starts an exercise program?

A 15. What is the average number of people who travel on I-40 between Raleigh and Durham each day?

D 16. Do Democrats tend to have lower salaries than Republicans?

C 17. In a typical day, how many more people visit Amazon.com than Target.com?

G 18. Do companies that advertise on more web sites tend to have higher profits?

G 19. Are cholesterol levels impacted by the number of servings of meat one eats per week?

D 20. Is the average final exam grade higher for classes with exams on the last day of the finals period than it is for those with exams on the first day?

Multiple Choice. (2 points each)

21. A 95% prediction interval based on $n = 20$ observations will be

a. narrower than a 95% CI for mean response based on the same data
b. narrower than a 90% prediction interval based on the same data
[✓ c.] wider than a 90% CI for mean response based on the same data

22. If X and Y are independent with standard normal distributions, Z = X +Y has mean

[✓ a.] 0
b. 1
c. 2

23. A multiple regression analysis with a value of R^2 near 0 indicates

a. that at least one of the explanatory variables are useful for predicting y.
b. that none of the explanatory variables are statistically significant
[✓ c.] that the explanatory variables explain little of the variation of y

24. The null hypothesis tested by the ANOVA F-test in 1-Way ANOVA is

a. All the slopes are zero
b. At least one mean is different from the others
[✓ c.] All treatments have the same mean

25. If $X \sim N(10,2)$, the standard deviation of a sample mean computed from a SRS of that population

a. Is greater than 2
[✓ b.] Is less than 2
c. We can't say- it depends on the sample size.

26. When doing an ANOVA analyses where we don't have any prior knowledge about which groups are likely to have the highest response values, we are most likely to use

a. Contrasts
[✓ b.] Multiple comparison procedures
c. Pooled t procedures

27. If we increase the sample size of a test using significance level 0.05, then

a. The probability of a Type I error will increase
b. The probability of a Type I error will decrease
[✓ c.] The probability of a Type I error will remain the same

28. The Law of Large Numbers tells us that

a. Large samples have normal distributions

[✓ b.] In large samples, the sample mean will be close to the population mean

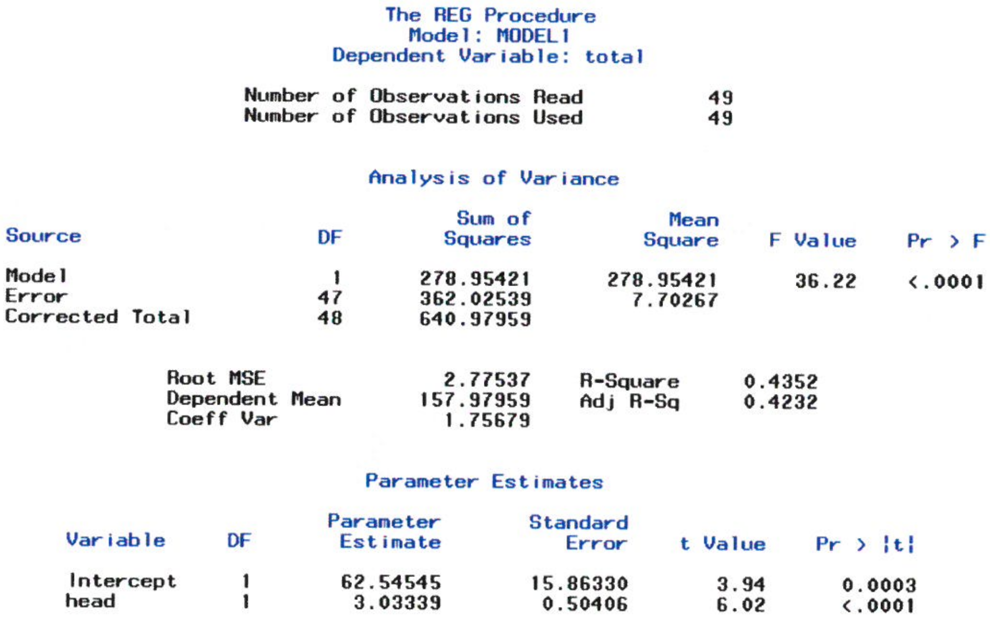
c. In large samples, estimates tend to be unbiased.

29. Stemplots would be most useful in conjunction with

[✓ a.] Analyses of a single population

b. Simple linear regression analyses

c. 1-Way ANOVA analyses



```
                              The REG Procedure
                                Model: MODEL1
                          Dependent Variable: total

                  Number of Observations Read          49
                  Number of Observations Used          49

                            Analysis of Variance

                                     Sum of           Mean
Source                   DF         Squares         Square    F Value    Pr > F

Model                     1       278.95421      278.95421      36.22    <.0001
Error                    47       362.02539        7.70267
Corrected Total          48       640.97959

            Root MSE              2.77537    R-Square     0.4352
            Dependent Mean      157.97959    Adj R-Sq     0.4232
            Coeff Var             1.75679

                            Parameter Estimates

                            Parameter       Standard
    Variable       DF        Estimate          Error    t Value    Pr > |t|

    Intercept       1        62.54545       15.86330       3.94      0.0003
    head            1         3.03339        0.50406       6.02      <.0001
```

30. In the SAS output above, the explanatory variable

[✓ a.] has a statistically significant effect on the response variable

b. explains almost all of the variation in the response variable

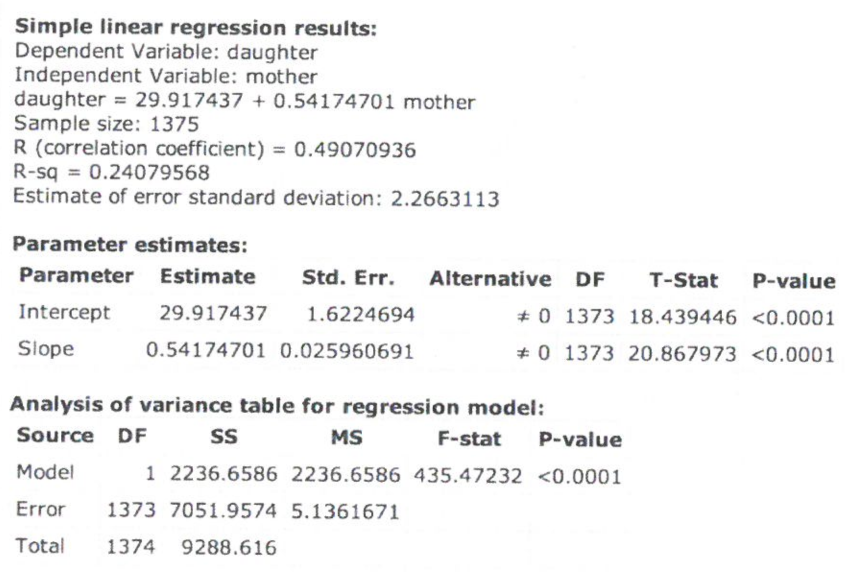
c. appears to have a nonlinear relationship with the response variable

31. In the SAS output above, the Total Mean Squares (MST) is

a. 286.66

[✓ b.] 13.35

c. 36.22



```
Simple linear regression results:
Dependent Variable: daughter
Independent Variable: mother
daughter = 29.917437 + 0.54174701 mother
Sample size: 1375
R (correlation coefficient) = 0.49070936
R-sq = 0.24079568
Estimate of error standard deviation: 2.2663113

Parameter estimates:
Parameter  Estimate     Std. Err.    Alternative  DF    T-Stat     P-value
Intercept  29.917437    1.6224694           ≠ 0  1373  18.439446  <0.0001
Slope      0.54174701 0.025960691           ≠ 0  1373  20.867973  <0.0001

Analysis of variance table for regression model:
Source  DF     SS         MS         F-stat     P-value
Model      1  2236.6586  2236.6586  435.47232  <0.0001
Error   1373  7051.9574  5.1361671
Total   1374   9288.616
```

32. The output above is from an analysis investigating the linear relationship between the heights of mothers and daughters. Use it to answer all of the following:

a) What was the name of the explanatory variable? (1 point) ~~Daughter~~ Mother

b) What was the name of the response variable? (1 point) Daughter

c) What was the statistical model used for this analysis (show the math, not just a name)? (2 points)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ , \quad \varepsilon_i \sim N(0, \sigma)$$

d) Provide numerical estimates for all three of the model parameters (2 points)

$\hat{\beta}_0 = 29.917$ $\quad \hat{\sigma} = 2.267$

$\hat{\beta}_1 = 0.5417$

e) What is the predicted value of the response variable when the explanatory variable has a value of 10? (2 points)

$$\hat{y} = 29.917 + 10(.5417)$$

f) Give a 95% confidence interval for the slope of the line relating these two variables. (3 points)

$$0.5417 \pm 1.96(.026)$$

↑

t w/ 1373 df ≈ z

33. The average number of total miles driven in 2008 for a SRS of 100 first-year drivers in NC was 5,200. Use this example to illustrate the difference between a population and a sample, and between a statistic and a parameter. (5 points)

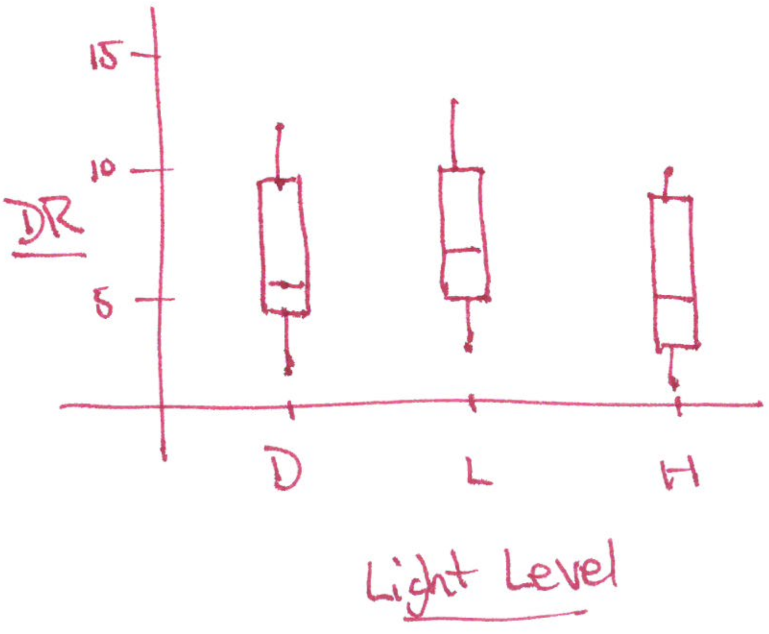
The population is all 1st-year drivers in NC, & they drive an avg of μ (unknown) miles per year; μ is a parameter

The sample of 100 1st-year drivers has an avg of 5,200 miles ($\bar{x}$); $\bar{x}$ is a statistic

34. The *diffusive resistance* (DR) of a leaf helps us understand how easily water can pass through the surface of that leaf. An experiment was carried out to investigate the effects of three light levels (Dark, Low, High) on DR. The results of that experiment are summarized in the following table; the entries in each cell are the **5-number summaries** for the DR values of light level:

	DR
Dark	2 4 6 9 12
Low	3 5 7 10 13
High	1 3 5 8 10

a. Draw and label an appropriate graph to help determine if there appears to be a difference in DR among the three light levels. (3 points)



(34-continued) Here is the ANOVA Table for analyzing the data above.

Source	df	SS	MS	F	P-value
Model	2	30	15	(III)	0.0015
Error	(I)	90	(II)		
Total	47	120			

b. What are the missing values labeled (I), (II), and (II) above? (1 point each)

I: 47-2=45 II: $\frac{90}{45} = 2$ III: $\frac{15}{2} = 7.5$

c. What is the *factor(s)* in this study? What are the *levels* of each factor? (2 points)

Light level is the factor. Its 3 levels are Dark, Low, & High

d. Are there statistically significant differences among the three light levels? Justify your answer. (2 points)

Yes. The ANOVA F-test has pvalue .0015

e. Do differences in light level do a good job of explaining the observed variation in dispersive resistance? Justify your answer (3 points)

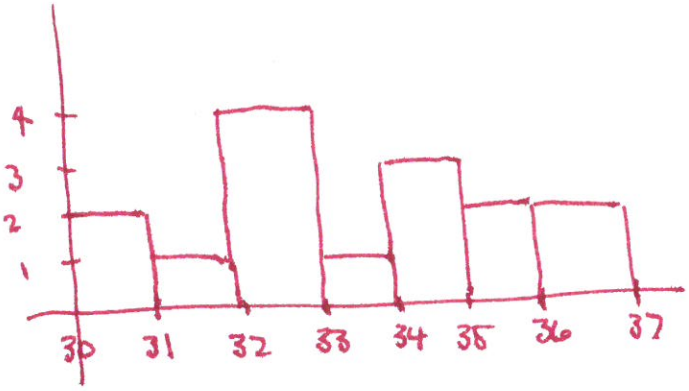
Not especially - the R^2 value is only $\frac{30}{120} = 0.25$.

35. The following data are 15 consecutive measures of monthly CO_2 levels at the Mauna Loa Observatory in Hawaii. The *sample standard deviation* for CO_2 level is 1.98.

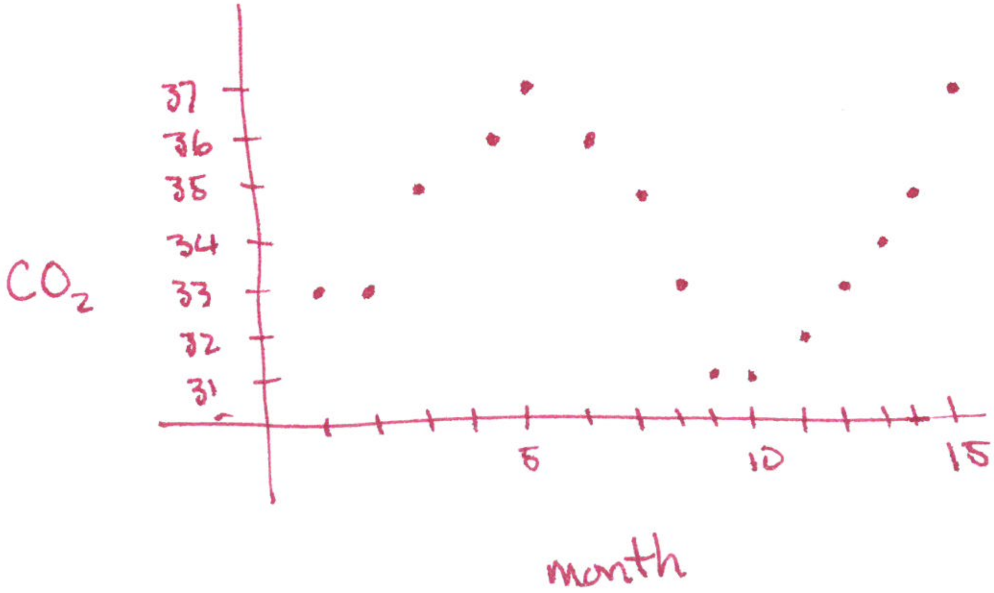
Month	CO_2		Month	CO_2		Month	CO_2	
1	33		6	36		11	32	
2	33		7	35		12	33	
3	35		8	33		13	34	
4	36		9	31		14	35	
5	37		10	31		15	37	

31 ||
32 |
33 ||||
34 |
35 |||
36 ||
37 ||

a. Display the distribution of CO_2 levels. (3 points)



b. Make a plot to look for a relationship between Month and CO_2 level. (3 points)



(35 continued)

c. Provide a numerical summary of the distribution of CO_2 levels. (2 points)

Since there is no evidence of strong skewness or outliers, use $\bar{x}$ for center & s for spread:

$$\bar{X} = \frac{548}{15} = 36.53$$

$$s = 1.98$$

Assume that CO_2 levels at Mauna Loa are known to follow a normal distribution.

d. Compute a 95% CI for the mean monthly CO_2 level. (3 points)

$$36.53 \pm t^*_{14}\left(1.98/\sqrt{15}\right)$$

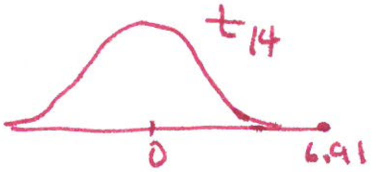
$$36.53 \pm 2.145\left(1.98/\sqrt{15}\right)$$

$$36.53 \pm 2.145\,(.511)$$

e. Is there strong evidence that the mean monthly CO_2 level is greater than 33? Justify your answer with a rigorous statistical calculation. (4 points)

$H_0: \mu = 33$

$H_A: \mu > 33$



↳ p-val < .005

$$t = \frac{36.53 - 33}{1.98/\sqrt{15}} = \frac{3.53}{.511} = 6.91$$

Since p-val << .01 ⇒ reject H_0 & conclude the mean CO_2 level is greater than 33.

36. Advil, Motrin, and Nuprin are all name brand versions of the drug Ibuprofen. Design an experiment using 120 subjects to determine if the three drugs are effective in reducing fever in children, and whether there are differences between the drugs. Explain what treatments you would use, how you would allocate subjects to treatments, and point out how you use good principles of experimental design. Finally, tell which statistical methods you would use to analyze the resulting data, and what specific questions they would address. (5 points)

We want to use the 3 key principles of design:

- control : include a placebo group
- randomize: randomly allocate 30 subjects to each group (Advil, Motrin, Nuprin, Placebo)
- replicate: by using 30 subjects/group, we have good replication.

To address the question of are there differences, we can look at the ANOVA F-test. We could then use contrasts (if we had planned comparisons), but more likely we would use a multiple comparison procedure to identify significant differences between pairs of the drugs (eg Advil vs Motrin, Motrin vs Nuprin, etc).