Additional Practice

*Also look into old exam, exam practice, homework, lecture examples, and discussion questions for final practice.
*Let me know via email or Piazza if you find any errors or things than need clarification ASAP so I can make any necessary updates.

1. In his book outliers, Malcolm Gladwell suggests that a hockey player's birth month has a big influence on his chance to make it to the highest levels of the game. Since January 1 is the cut-off day for youth leagues in Canada, January-born players will be up to one year older than those they are competing against and thus be bigger, stronger, and get more playing time, which results in a better chance of being successful. A random sample of 80 NHL players from a recent season was selected and the month of their birthdays were recorded. The data is given below. Do these data give convincing evidence that the birthdays of NHL players are not uniforly distributed throughout the four quarters of the year? Conduct an appropriate hypothesis test.

Month	Jan-March	April-June	July-Sept	Oct-Dec
Number of Players	32	20	16	12

2. An experiment was conducted to determine the concentration of a particular bacterium (*Pseudomonas syringae*) found adhering to rocks in river beds. Of particular interest was whether the number of bacteria was the same for rocks near the source of the river versus rocks near the outlet of the river. The experiment was conducted as follows. Six rivers in Iowa were randomly sampled. Then, for each river, the number of bacteria was measured for rocks at the source of the river and for rocks at the outlet of the river (and measured as number of bacteria per unit sample). The data are provided below with some summary statistics:

River	1	2	3	4	5	6	Sample Mean	Sample Variance
Source	5600	2600	3260	4910	3750	1720	3640	2075800
Outlet	5480	2380	3300	4800	3680	1600	3540	2107520

   (a) Construct a 99% confidence interval for the difference in the mean number of bacteria at the source compared to the mean number of bacteria at the outlet. (Note: the experimenters are quite confident that the variance of the number of bacteria is the same at the source as at the outlet. They also believe that the data are normally distributed.)

   (b) Without doing any further work, comment on what conclusions you could draw if you conducted a test of the null hypothesis that the mean number of bacteria is the same at the source and at the outlet, versus the two-sided alternative.

3. A study will be conducted to determine the impact of a generating station in Portage, Wisconsin, on air quality. SO_2 concentration is to be measured at several sites at similar distances from the power plant. After the data is collected, a two-sided test will be used at level $\alpha = 0.05$ in order to test the null hypothesis that the mean SO_2 concentration in the air is equal to 10 μg SO_2/m^3 versus an alternative of a difference. This is thought to be an acceptable SO_2 concentration. The study aims at verifying that the power plant is still "clean" after several years of operation.

   (a) A mean value of 30 μg SO_2/m^3 is thought to be unacceptable. It is desired that such a meam level of SO_2 pollution have a 90% chance of being detected by the experiment. How many sites need to be sampled in order to achieve this goal? Note: Previous data collected near other power plants have shown that the standard deviation in SO_2 concentration is 18 μg/m^3 (variance of 18^2) and that a normality assumption is reasonable.

(b) If the test is to be done at level $\alpha = 0.01$, and if it is still desired that a level of 30 μg SO_2/m^3 be detected with a 90% chance, should one sample:

□ fewer, □ the same number of or □ more sites

than in 3a? Check the appropriate box and explain concisely. (Hint: No formal calculations are needed.)

(c) If the test is made at level $\alpha = 0.05$ as in 3a, but if the number of sites sampled is four times larger than in 3a, which level of SO_2 pollution will have a 90% chance of being detected:

□ 7.5 □ 15 or □ 20 μg SO_2/m^3

Check the appropriate box and explain concisely. (Hint: There is no need for more calculations than using the formula for the sample size.)

4. A study was conducted to investigate the sweetness of juice obtained from three different varieties of grape used to make wine. To do the experiment, 15 plots were located at random in a winery, and at random, 5 plots were planted with the variety NorthStar, 5 plots were planted with the variety SweetCab, and 5 plots were planted with ZinnRed. At the end of the growing season, grapes were sampled from each plot and the sweetness of the grapes was measured (in degrees Brix, a standard used in wine making).

The data are as follows:

NorthStar 21.3, 22.7, 19.1, 19.6, 20.0

SweetCab 28.1, 24.6, 26.1, 23.3, 22.9

ZinnRed 20.3, 26.0, 24.2, 19.8, 21.2

Here are some summary results:

Variety	Sample Mean	Sample Standard Deviation
NorthStar	20.54	1.46
SweetCab	25.00	2.14
ZinnRed	22.30	2.68

(a) Complete the following Analysis of Variance table:

Source	df	SS	MS
Variety			
Error			
Total			

(b) Carry out a test for equality of sweetness among the three varieties aat a 5% significance level. State the null and alternative hypotheses. Interpret the resulting p-value. (You may assume that the appropriate assumptions are met, without checking them.)

(c) State the assumptions underlying this analysis. (You are not required to assess these assumptions.)

(d) Do a pairwise comparison of the three groups using confidence intervals and no multiple comparison correction (a Fischer LSD proceedure) and summarize your findings in a table.

5. In a study of plant disease epidemiology, researchers inoculated several randomly sampled potato plants with a pathogen and then recorded how long it took, in days, before each of the plants exhibited disease symptoms. The researchers were particularly interested in comparing these times for two different varieties of potato: Russet Burbank (RB) and Yukon Gold (YG). The data are given below:

RB	16	14	13	18	10	12	13	15	19	12	11
YG	24	57	17	84							

Here are some summary statistics for these data:

Variety	sample mean	sample variance
RB	13.91	8.09
YG	45.50	963.00

(a) The investigators in this experiment begin by assuming that the two groups have equal variance. They are also willing to assume any necessary normality.

Based on those assumptions, perform a formal test to assess whether there is evidence that the mean time for disease symptoms to develop is different for the two varieties, versus a null of equality.

(b) Upon further thought, the investigators decided that they are unwilling to assume that the two groups have equal variance. However, they are still willing to assume any necessary normality. Therefore, they proceed as follows:

Complete this test, including p-value and interpretation.

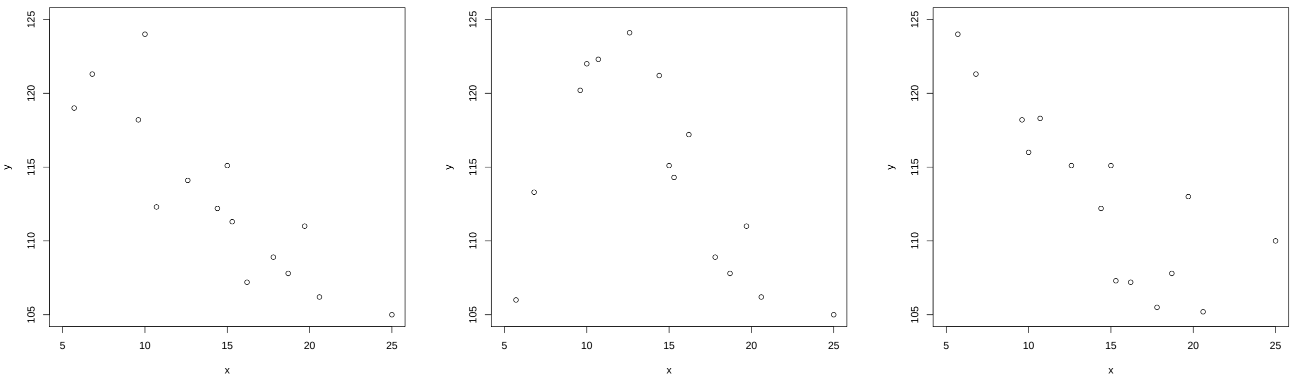
(c) Upon further reflection, the investigators have decided that they are also uncomfortable with the assumption of normality.

State an appropriate nonparametric test to use in this situation. (You are not required to complete the test.)

6. Variations in clay brick masonry weight have implications not only for structural and acoustical design but also for design of heating, ventilating, and air conditioning systems. The article "Clay Brick Masonry Weight Variation" (*J. of Architectural Engr.* 1996: 135-137) gave a scatter plot of y = mortar dry density (lb/ft^3) versus mortar air content (%) for a sample of mortar specimens, from which the following representative data was read:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x	5.7	6.8	9.6	10.0	10.7	12.6	14.4	15.0	15.3	16.2	17.8	18.7	19.7	20.6	25.0
y	119.0	121.3	118.2	124.0	112.3	114.1	112.2	115.1	111.3	107.2	108.9	107.8	111.0	106.2	105.0

A linear model was fit to the data in the table above as all necessary assumptions seemed to be satisfied.Given that $\bar{x}_i = 14.54$, SSErr = 112.443 and $\sum_{i=1}^{15}(x_i - \bar{x})^2 = 405.836$. It was found that $\hat{\beta_1} = -.92$ **and** $\hat{\beta_0} = 126.25$



(a) In the scatter plots above, the given data is on the left, and two other datasets make up the two scatter plots to the right. For which scatter plot does a homoscedastic simple linear regression model seems most appropriate?

(b) Calculate a 95% confidence interval for $\hat{\beta}_1$. From the result of CI, what conclusion can you draw about the relationship between x and Y? Is it consistent with the scatter plot?

(c) Predict the value of Y at $x^* = 11$ by using the regression model. Find 95% prediction interval Y at $x^* = 11$.

(d) Which prediction interval will be wider, the one at $x^* = 5.9$ or the one at $x^* = 14.6$? Explain without calculations.

7. Suez et al. (*Nature* 2014, 514:181) studied the effect of artificial sweeteners on blood glucose regulation. In one experiment, they fed mice with a normal diet, supplemented in their drinks either by glucose or by saccharin (a non-caloric artificial sweetener). Mice were randomly selected to receive either the glucose or the saccharin treatment. After 22 weeks, the mice were subject to a glucose tolerance test to measure their ability to regulate glucose in their blood. A high glycemic response means poor regulation (high glucose concentrations). Part of their glycemic response data is given below for 11 mice in each treatment.

glucose	22.1	13.6	16.3	17.6	14.3	16.6	17.0	18.8	15.2	14.9	15.6
saccharin	21.8	22.3	16.8	31.2	20.6	19.3	27.4	20.9	23.2	22.1	18.6

(a) State (briefly) the assumptions you must make to proceed with an analysis of this problem. Define all terms. (You do not need to assess the validity of the assumptions for this question.)

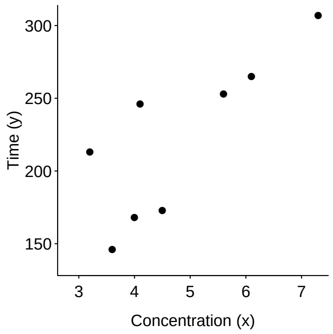
(b) Using the data above, perform a hypothesis test of the claim that mice under the two supplement treatments (glucose and saccharin) have the same mean glycemic response, versus the two-sided alternative.

(c) Compute a 95% CI for the difference in mean glycemic response in mice between the two treatments.

(d) Test the hypothesis that the mean glycemic response in mice getting a saccharine supplement equals the mean glycemic response in mice getting a glucose supplement plus 2.5 (versus the 2-sided alternative).

8. An experiment was designed to study the relationship between the initial concentration of bacteria in a test tube, and the time it takes for the number of bacteria in the test tube to grow so large that you cannot see through the test tube. The experiment was conducted as follows: 8 test tubes were each prepared with a bacterial suspension; the concentration (x) was different for each test tube. Each tube was monitored, and the experimenter measured the time (Y) it took until she could no longer see through the tube (i.e. the tube became opaque).

The data are plotted below.



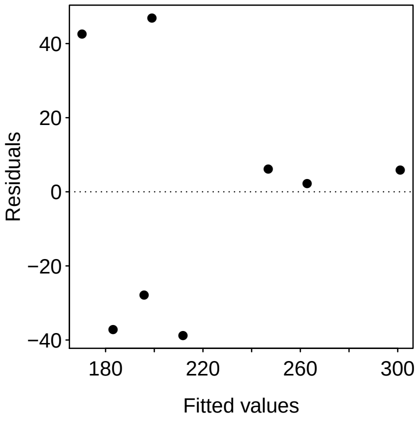
Some summary values are:

$$\begin{aligned}
\bar{x} &= 4.8 \\
\bar{y} &= 221.375 \\
\sum_{i=1}^{8}(x_i-\bar{x})(y_i-\bar{y}) &= 439.9 \\
\sum_{i=1}^{8}(x_i-\bar{x})^2 &= 13.8 \\
\sum_{i=1}^{8}(y_i-\bar{y})^2 &= 21781.875 \\
\text{SSErr} &= 7759.3
\end{aligned}$$

(a) Find the least squares estimates of the slope and intercept for the regression of Y on x.

(b) Is there evidence that the slope differs from 60 at level 0.05? Perform necessary test.

(c) In the data set, the first tube had an observed concentration of 4.1 and an observed time of 246. In the residual plot below, circle the residual that corresponds to this tube. Justify your choice.



9. Observations of 80 litters, each containing 3 rabbits, reveal the following frequency distribution of the number of male rabbits per litter.

Number of males in litter	0	1	2	3	Total
Number of litters	19	32	22	7	80

Under the model of Bernoulli trials for the sex of rabbit, the probability distribution of the number ofmales per litter should be binomial with 3 trials and p=probability of a male birth. From these data, the parameter p is estimated as

$$\hat{p} = \frac{Totalnumberofmalesin80litters}{Totalnumberofrabbitsin80litters} = \frac{97}{240} \approx 0.4$$

.

(a) Using the binomial model for the three trials and $p = 0.4$, determine the expected cell probabilities.

(b) Perform an appropriate test to determine if the observations from the 80 litters of 3 rabbits look to be consistent with the Bernoulli model with p=0.4.

10. To compare the effectiveness of four drugs in relieving postoperative pain, an experiment was done by randomly assigning 195 surgical patients to the drugs under study. Recorded here are the number of patients assigned to each drug and the number of patients who were free of pain for a period of five hours.

	Free of Pain	No of Patients assigned
Drug 1	23	53
Drug 2	30	47
Drug 3	19	51
Drug 4	29	44

(a) Make a contingency table showing the counts of patients who were free of pain and those who had pain, and test the null hypothesis that all four drugs are equally effective (use $\alpha = 0.05$).

(b) Let p_1, p_2, p_3, and p_4 denote the population proportions of patients who would be free of pain under the use of drugs 1, 2, 3, and 4, respectively. Calculate a 90% confidence interval for p_1 and p_2.

(c) Make a 2X2 contingency table and test $H_o : p_1 = p_3$ versus $H_o : p_1 \neq p_3$ at $\alpha = 0.05$ employing (i) the χ^2 test, and then (ii) the Z test. Make sure to draw conclusions in context.

(d) Construct and interpret a 95% confidence interval for the difference $p_4 - p_2$.