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

ANSWER: H_o*: the birthdays of all NHL players are evenly distributed across the 4 quarters of the year, ie* $\pi_{Q1} = \pi_{Q2} = \pi_{Q3} = \pi_{Q4} = 0.25$*; and* H_A*: the birthdays are not evenly distributed across the 4 quarters. The data come from a random sample of a large population of NHL players so independence of observations is reasonable. We will check that expected count values after their computation. Expected counts for each quarter is* $80 * .25 = 20$*. Since these are all larger than 5, our chi-squared distribution is sufficient for calculating a p value from an observed* χ^2*.* $\chi^2_{obs} = \frac{(32-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(12-20)^2}{20}$*=11.2. pvalue=*$P(\chi^2_{df=3} \geq 11.2) = 0.011$ *or from a table, between 0.01 and 0.02. Since the p value is less than* $\alpha = 0.05$*, we have sufficient evidence to reject* H_0*. We have convincing evidence that the birthdays of NHL players are not evenly distributed across the four quarters of the year.*

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.)

   ANSWER: These are paired samples since measurements are gathered from two different places of same rivers. As such, the differences need to be calculated. Taking the source of the river minus the outlet of the river, the mean difference ($\bar{x}_D$) *is* 100 *and the sample standard deviation of the differences* (s_D)

is 84.617. *Therefore, 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 is*

$$\bar{x}_D \pm t_{(5,0.01/2)} * \frac{s_D}{\sqrt{n}} = 100 \pm 4.032 \left(\frac{84.617}{\sqrt{6}} \right) = 100 \pm 139.2841 = (-39.284, 239.284)$$

(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.

ANSWER: Because the confidence interval in part (a) contains zero, we would fail to reject the null hypothesis at level 0.01.

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 $\mu g/m^3$ (variance of 18^2) and that a normality assumption is reasonable.

ANSWER: $n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{(\mu_0 - \mu_1)^2} = \frac{(z_{0.025} + z_{0.1})^2 \sigma^2}{(\mu_0 - \mu_1)^2} = \frac{(1.96 + 1.28)^2 (18)^2}{(10 - 30)^2} = 8.505$. *Hence required sample size is* $n = 9$ *to achieve* 90% *power for detecting the* SO_2 *concentration of 30* $\mu g/m^3$.

(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.) *ANSWER: If we decrease the level of significance to* $\alpha = 0.01$, *the power also decreases if the sample size stays the same. To compensate and maintain the power to the desired 90% when* $\mu_A = 30$, *one should sample more sites.*

(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.)

ANSWER: With more samples, a smaller difference from the null can be detected. More specifically, *the new* μ_1 *that can be detected with 90% power satisfies:*

$$\begin{aligned} 4n = 4\,\frac{(1.28+1.96)^2 * (18)^2}{(10-30)^2} &= \frac{(1.28+1.96)^2 * (18)^2}{(10-\mu_1)^2} \\ \frac{4}{20^2} &= \frac{1}{(10-\mu_1)^2} \\ |10-\mu_1| &= \frac{1}{\sqrt{4}} \times 20 = 10 \\ \mu_1 &= 20 \ \ (or\ 0) \end{aligned}$$

Thus, $\mu_1 = 20$.

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			

ANSWER: $MSEr = ((5-1)\times 1.46^2 + (5-1)\times 2.14^2 + (5-1)\times 2.68^2)/(15-3) = 4.631$. *Overall mean:* $\frac{20.54*5+25*5+22.30*5}{15} = 22.61333$ $SSVariety = 5*(20.54-22.61333)^2 + 5*(25-22.61333)^2 + 5*(22.30-22.61333)^2 = 50.46533$

Source	df	SS	MS	F	pvalue
Variety	2	50.465	25.233	25.233/4.631=5.4487	0.0207191
Error	12	55.574	4.631		
Total	14	106.039			

(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.)

ANSWER: $F = 25.211/4.631 = 5.44$ *on 2 and 12 df, hence the p-value is* $P(F_{2,12} > 5.44)$ *which is less than 0.05 and hence we will reject* H_0 *at level 0.05.*

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

ANSWER: (1) Independence between groups, and independence between all observations with each group; (2) equal population variance across all the groups; (3) population of errors are normally distributed and centered at zero $\epsilon iid \sim N(0, \sigma^2)$.

(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.

ANSWER

$\mu_{NS} - \mu_{SC} : 20.54 - 25.00 \pm t_{0.025,12}\sqrt{4.631(1/5 + 1/5)} = -4.46 \pm 2.18 * 1.36 = (-7.42, -1.50)$

$\mu_{NS} - \mu_{ZR} : 20.54 - 22.30 \pm t_{0.025,12}\sqrt{4.631(1/5 + 1/5)} = -1.76 \pm 2.18 * 1.36 = (-4.72, 1.2048)$

$\mu_{SC} - \mu_{ZR} : 25.00 - 22.30 \pm t_{0.025,12}\sqrt{4.631(1/5 + 1/5)} = 2.7 \pm 2.18 * 1.36 = (-0.26, 5.66)$

Variety	Mean	Group
NorthStar	20.54	A
ZinnRed	22.30	AB
SweetCab	25.00	B

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.

ANSWER: First, $s_p^2 = ((11 - 1) \times 8.09 + (4 - 1) \times 963.00)/(11 + 4 - 2) = 228.45$. *Thus* $T_{obs} = \dfrac{13.91 - 45.50}{\sqrt{228.45}\sqrt{\frac{1}{11} + \frac{1}{4}}} = -3.58.$

P-value is $2 \times P(T_{13} \geq |-3.58|) < 0.01$. *Thus the researchers can conclude that on average there is a difference in these two types of potatoes at level 0.01.*

(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:

$$\tilde{T} = \frac{13.91 - 45.50}{\sqrt{\frac{8.09}{11} + \frac{963.00}{4}}} = -2.03$$

Complete this test, including p-value and interpretation.

ANSWER: We have $r_1 = 8.09/11 = 0.74$ *and* $r_2 = 963.00/4 = 240.75$. *Hence* $df = m = \dfrac{(r_1+r_2)^2}{\dfrac{r_1^2}{n_1-1} + \dfrac{r_2^2}{n_2-1}} =$ 3.02. *We round down and use* T_3 *to obtain a p-value between 0.10 and 0.20. For this test there is no evidence of a difference between the two varieties of potato at level 0.05.*

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

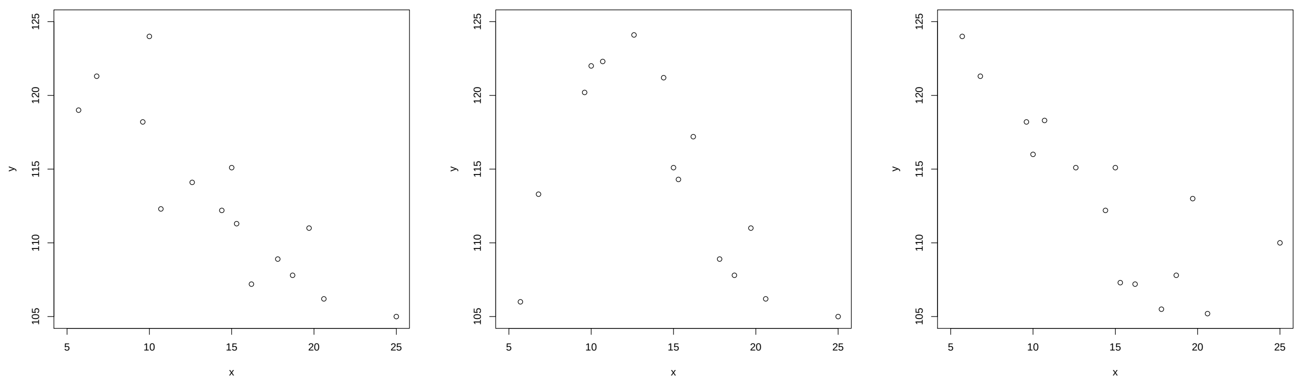
State an appropriate nonparametric test to use in this situation, and calculate the appropriate test statistic. (You are not required to complete the test.)

ANSWER: Use a Wilcoxon rank sum 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?

ANSWER: Homoscedastic means that the errors have equal population variance and heteroscedastic means that the errors do not have equal population variance.

__Left__: A straight line would make a good sense for this plot, there seems to be a linear relationship between the variables. The random error around a fitted line would likely to have constant variance and hence a homoscedastic linear regression model would be appropriated here.
__Center__: The relationship between the variables does not seem to be linear. So a simple linear regression model would not be appropriate.
__Right__: Although the relationship between the variables looks to be approximately linear, the random error around the fitted line likely would fail the constant variance assumption (note the fanning pattern). It seems that as fitted values grow larger so too would the residual. Hence a heteroscedastic linear regression model would be appropriated here.

(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?

ANSWER: 95% *CI of* β_1 *is* $\hat{\beta}_1 \pm t_{(13,0.025)} * \hat{SE}(\hat{\beta}_1)$ *where* $\hat{SE}(\hat{\beta}_1) = \dfrac{\sqrt{MSErr}}{\sqrt{\sum_{i=1}^{15}(x_i-\bar{x})^2}} = \dfrac{\sqrt{\frac{SSErr}{15-2}}}{\sqrt{\sum_{i=1}^{15}(x_i-\bar{x})^2}}$.

Now, $\hat{\beta}_1 = -0.92$, $\hat{SE}(\hat{\beta}_1) = \dfrac{\sqrt{\frac{112.443}{13}}}{\sqrt{405.836}} = 0.1460$ *and* $t_{(13,0.025)} = 2.16$. *Therefore the* 95% *CI of* β_1 *is* $(-0.92 \pm 2.16 * 0.1460) = (-1.2353, -0.6047)$. *Hence we can reject* $H_0 : \beta_1 = 0$ *at level 0.05, i.e., we can conclude that there is a linear relationship between* x *and* Y. *It is consistent with the scatter plot as scatter plot is also showing a strong linear relationship between* x *and* Y.

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

ANSWER: $(\hat{Y}|x^* = 11) = 126.25 - 0.92(11) = 116.13$. 95% *prediction interval of* Y *at* $x^* = 11$ *is* $(\hat{Y}|x = 11) \pm t_{(13,0.025)} * \left[\sqrt{SSErr/(n-2)} * \sqrt{1 + \frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_{i=1}^{11}(x_i-\bar{x})^2}}\right]$, *i.e.,* $(116.13 \pm 2.16 * 3.0810) = (109.4750, 122.7850)$.

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

ANSWER: A prediction for a value of $x^* = 5.9$ *would have a larger standard error because 5.9 is much farther from* $\bar{x}$ *than 14.54 and thus yielding a larger standard error and wider prediction interval.*

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. Here the two samples are independent, not paired. Let $\{X_1, \ldots, X_{n_1}\}$ denotes the glycemic response (in mg/dl) of the mice that received glucose treatment, and the sample given in this problem is

$\{22.1, 13.6, 16.3\ , 17.6, 14.3, 16.6, 17.0, 18.8, 15.2, 14.9, 15.6\}$. Let $\{Y_1, \ldots, Y_{n_2}\}$ denotes the glycemic response (in mg/dl) of the mice that received saccharin treatment, and the sample given in this problem is $\{21.8, 22.3, 16.8, 31.2, 20.6, 19.3\ , 27.4, 20.9, 23.2, 22.1, 18.6\}$.

i. In order to do the two-sample T-test, the assumptions should be satisfied first that $\{X_1, \ldots, X_{n_1}\}$ is a random sample from $N(\mu_1, \sigma_1^2)$, and $\{Y_1, \ldots, Y_{n_2}\}$ is a random sample from $N(\mu_2, \sigma_2^2)$, $\{X_1, \ldots, X_{n_1}\}$ and $\{Y_1, \ldots, Y_{n_2}\}$ are independent of each other.

ii. $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 \neq 0$.

- From the data, $n_1 = n_2 = 11$, our observed $\bar{x} = 16.55$, $\bar{y} = 22.20$, $s_1^2 = 5.64$, $s_2^2 = 16.4$. Hence $s_2^2/s_1^2 = 16.4/5.64 = 2.91 < 4$. Hence we can assume $\sigma_1^2 = \sigma_2^2$ based on thumb rule. Therefore the test statistic T is $\dfrac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$. Under H_0, T will follow $T_{n_1+n_2-2}$. Now $s_p = \sqrt{\dfrac{(11-1)5.64 + (11-1)16.4}{11 + 11 - 2}} = 3.32$, $T_{obs} = \dfrac{16.55 - 22.2}{3.32\sqrt{\frac{1}{11} + \frac{1}{11}}} = -3.99$. Two sided p-value is $2P(T_{20} > 3.99) < 2 \times 0.001 = 0.002$, which implies that we can reject H_0 i.e. we can conclude that mice under the two supplement treatments (glucose and saccharin) have different mean glycemic response.
- (**Just to show the Welch's t-test procedure**). If we could not assume equal population variances, we have to use the Welch t-test. Test Statistic is $T = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$, under H_0, it follows T-distribution with an adjusted degree of freedom expressed by $\dfrac{(r_1 + r_2)^2}{\frac{r_1^2}{n_1-1} + \frac{r_2^2}{n_2-1}}$, where $r_1 = s_1^2/n_1 = 5.64/11 = 0.51, r_2 = s_2^2/n_2 = 16.4/11 = 1.49$.
  In this case, adf $= \dfrac{(0.51 + 1.49)^2}{\frac{0.51^2}{11-1} + \frac{1.49^2}{11-1}} = 16.1$, which is rounded down to 16. $T_{obs} = \dfrac{16.55 - 22.2}{\sqrt{\frac{5.64}{11} + \frac{16.4}{11}}} = -3.99$, p-value is $2P(T_{16} > 3.99) < 2 \times 0.001 = 0.002$. Therefore we can conclude that mice under the two supplement treatments (glucose and saccharin) have different mean glycemic response.

(c) Compute a 95% CI for the difference in mean glycemic response in mice between the two treatments.
ANSWER: A $100(1-\alpha)\%$ *CI for the difference in mean is*

$$(\bar{x} - \bar{y}) \pm T_{(n_1+n_2-2,\alpha/2)} * s_p * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

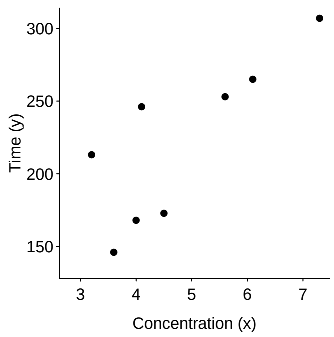
where $T_{(20,.025)} = 2.086$. *Hence* 95% *CI of the difference of mean glycemic response is* $\Big((16.55 - 22.2) - 2.086 \times 3.32\sqrt{\frac{1}{11} + \frac{1}{11}}, (16.55 - 22.2) + 2.086 \times 3.32\sqrt{\frac{1}{11} + \frac{1}{11}}\Big)$ *mg/dl, i.e.* $(-8.65, -2.75)$ *mg/dl.*

(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).
To test $H_0 : \mu_1 - \mu_2 = -2.5$ *versus* $H_1 : \mu_1 - \mu_2 \neq -2.5$. *The test statistic* T *is* $\dfrac{\bar{X} - \bar{Y} - (-2.5)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$. *Under* H_0, T *will continue to follow* T_{20}, *and the observed test statistic* $T_{obs} = \dfrac{16.55 - 22.2 - (-2.5)}{3.32\sqrt{\frac{1}{11} + \frac{1}{11}}} = -2.225$,
The p-value is $2P(T_{20} > 2.225) < 2P(T_{20} > 2.086) = 0.05$ $(T_{(20,0.025)} = 2.086)$. *Therefore at level 0.05*

we can reject the claim 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.

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:

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

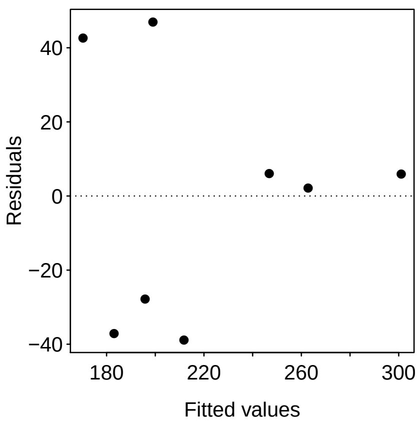
   (a) Find the least squares estimates of the slope and intercept for the regression of Y on x.

   ANSWER: $\hat{\beta}_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{439.9}{13.8} = 31.877$. *The LSE of intercept is* $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 221.375 - 31.877 * 4.8 = 68.366$

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

   ANSWER: $T_{obs} = \frac{\hat{\beta}_1 - 60}{\sqrt{MSErr/\sum_{i=1}^{8}(x_i-\bar{x})^2}} = \frac{31.877-60}{\sqrt{1293.217/13.8}} = -2.905$. *From t-Table,* $P(T_6 > 2.447) = 0.025$, *so the p-value* < 0.05 *and hence we can conclude at level 0.05 that the slope can not be assumed to be 60.*

   (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.



ANSWER: The fitted value for this observation is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 68.366 + 31.877 * (4.1) = 199.06$ *and the corresponding residual is* $r = y - \hat{y} = 246 - 199.06 = 46.94$*. So, the point on the residual plot is the top most point.*

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.
ANSWER: We need to calculate P(M=0), P(M=1), P(M=2), and P(M=3) under a binomial assumpgion where $M\ Bin(3, .4)$.
$P(M = 0) = .4^0 * .6^3 = 0.216$; $P(M = 1) = \binom{3}{1}.4^1 * .6^2 = 0.432$; $P(M = 2) = \binom{3}{2}.4^2 * .6^1 = 0.288$; $P(M = 3) = \binom{3}{3}.4^3 * .6^0 = 0.064$

(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.

Number of males in litter	0	1	2	3	Total
Observed Number of litters	19	32	22	7	80
Expected Number of litters	80*.216=17.28	80*.432=34.56	80*.288=23.04	80*.064=5.12	80
$(O-E)^2/E$	$\frac{(19-17.28)^2}{17.28}$	$\frac{(32-34.56)^2}{34.56}$	$\frac{(22-23.04)^2}{23.04}$	$\frac{(7-5.12)^2}{5.12}$	
χ^2 value:	0.1712037	0.1896296	0.04694444	0.6903125	1.09809

pvaluee: $P(\chi^2_3 > 1.09809) = 0.7775352$. The probability of seeing this amount of variability in a sample from a population that truely matches a bournoulli model with p=0.4 is very high. Therefore we have no evidence against the null of a bournoulli probability model with p=0.4 describing the number of male rabbits in litter of size 3.

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 $\alpha = 0.05$).

	Obs Free of Pain	Obs Has Pain	No of Patients assigned
Drug 1	23	30	53
Drug 2	30	17	47
Drug 3	19	32	51
Drug 4	29	15	44
Total	101	94	195

$H_o : p_{y1} = p_{y2} = p_{y3} = p_{y4}$ and $p_{n1} = p_{n2} = p_{n3} = p_{n4}$. Vs H_a : the pain distribution is different for at least one drug (from the others).

	Exp Free of Pain	Exp Has Pain	No of Patients assigned
Drug 1	27.45128	25.54872	53
Drug 2	24.34359	22.65641	47
Drug 3	26.41538	24.58462	51
Drug 4	22.78974	21.21026	44

$\chi^2 = 0.7218 + ... + 1.81833 = 12.05279$

pvalue: $P(\chi^2_3 > 12.05279) = 0.007204$. We have strong evidence against the null. Evidence suggests not all drugs have the same pain-free rate.

(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.

ANSWER:

For p_1: $\hat{p_1} \pm z_{.05}\sqrt{\frac{(\hat{p_1})(1-\hat{p_1})}{n_1}}$ so $23/53 \pm 1.645 * \sqrt{\frac{(23/53)(30/53)}{53}} = 0.4339623 \pm 1.645 * 0.06807862 = (0.322, 0.546)$

For p_2: $\hat{p_2} \pm z_{.05}\sqrt{\frac{(\hat{p_2})(1-\hat{p_2})}{n_2}}$ so $30/47 \pm 1.645 * \sqrt{\frac{(30/47)(17/47)}{47}} = 0.638 \pm 1.645 * 0.07008713 = (0.523, 0.753)$

(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.

Drug	Free	Pain	Total
Drug 1	23	30	53
Drug 3	19	32	51
Total	42	62	104

Using the χ^2 test:

Expected Counts:

Drug	Free	Pain	Total
Drug 1	21.40385	31.59615	53
Drug 3	20.59615	30.40385	51
Total	42	62	104

χ^2 statistic: 0.11903+0.0806+0.1237+0.0838=0.40716. p value: $P(\chi^2_1 > 0.40716) = 0.5234$

Using the Z test:

$H_o: p_1 = p_3$ or $H_o: p_1 - p_3 = 0$. We calculate $p_c = \frac{23+19}{53+51} = 42/104$.

$z = \frac{(\hat{p_3} - \hat{p_1}) - 0}{\sqrt{\frac{(42/104)(62/104)}{53} + \frac{(42/104)(62/104)}{51}}} = \frac{0.06141324}{0.09624558} = 00.6380889$. (Notice $0.6380889^2 = 0.4071574$). pvalue$=2 * P(Z > 0.6381) = 2 * 0.2617043 = 0.5234086$, same as we got above.

We have no evidence of a difference in the proportion of patients who will be pain free with Drug 1 and Drug3.

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

95% CI: $\hat{p_4} - \hat{p_2} \pm z_{.975} * \sqrt{\frac{\hat{p_4}(1-\hat{p_4})}{n_4} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}} = 29/44 - 30/47 \pm 1.96\sqrt{\frac{(29/44(1-29/44)}{44} + \frac{(30/47(1-30/47)}{47}} = 0.02079304 \pm 1.96 * 0.100094 = (-0.1753, 0.2169)$.We are 95% confident this interval captures the true difference in rates of Pain Free for the patients on drug 4 and drug 2. Since this confidence interval for the differnce in success rate contains 0, we have no evidence that drug 4 or drug 2 have different efficacy in treating patient pain.