Stat 101 HW#10

1. **Titanic** Here is a table showing who survived the sinking of the *Titanic* based on whether they were crew members, or passengers booked in first, second, or third class staterooms:

	Crew	First	Second	Third	Total
Alive	212	202	118	178	710
Dead	673	123	167	528	1491
Total	885	325	285	706	2201

Let's see if someone's chances of surviving were the same regardless of their status on the ship.

a) State the null and alternative hypotheses we would test here.
b) How many degrees of freedom would the statistic have?

Let's see how to do Problem in JMP. The main idea is that we have to have three columns: One for Class, one for Survival and one for the number of people in each category. So it looks like this:

Survive	Class	Counts
Alive	Crew	212
Alive	First	202
Alive	Second	118
Alive	Third	178
Dead	Crew	673
Dead	First	123
Dead	Second	167
Dead	Third	528

From this, you can do the Chi-square test from **Fit Y by X**, where **Survive** is Y, **Class** is X and **Counts** is **Freq**.

c) Show the Contingency table.
d) Show where the degrees of freedom and the Chi-square statistic are located.
e) Double click on the p-value. Change the Format from p-value to Fixed Dec and select 10. What is the actual p-value?

2. **Survival and Gender** Newspaper headlines at the time, and traditional wisdom in the succeeding decades, have held that women and children escaped the *Titanic* in greater proportion than men. Here is a table with data by gender. Do you think that survival was independent of gender? Defend your conclusion.

	Female	Male	Total
Alive	343	367	710
Dead	127	1364	1491
Total	470	1731	2201

3. **Survival and Gender, One More Time** In Exercise 2 you could have checked for a difference in the chances of survival for men and women using two-proportion z-procedures (chapter 22). For a two by two table, testing proportions using *z* and testing homogeneity using χ^2 are equivalent. The χ^2 test generalizes to more groups, though.
   a) Find the *z* –value for this approach.
   b) Show that the square of your calculated value of z is the value of χ^2 you calculated in Exercise 2.
   c) Show that the resulting P-values are the same.

4. **Patterns of Pi.** We all know that π=3.1415926… Are the digits random, or do they contain some mysterious hidden pattern? In a recent Math colloquium, a student tested uniformity on the first 1,000,000 digits of π and found the following:

Digit	Count
0	99959
1	99758
2	100026
3	100229
4	100230
5	100359
6	99548
7	99800
8	99985
9	100106
Total	1000000

   a) To test whether this is uniform, would this be a test of homogeneity, independence or goodness-of-fit?
   b) How many degrees of freedom does it have?
   c) Test the appropriate null hypothesis at α=.05.
   (To do it in JMP, put the 10 digits in a column and the counts in another column. Go to **Distribution,** putting in the digits as the variable and counts as **Freq**. Now, in **Distribution,** find **Test Probabilities** under the red triangle. Put in the appropriate null hypothesis probabilities (from the uniform). Click Done. Find the Chi-square statistic and its p-value.)

5. **Wine Prices.** The data set Winery Prices contains the average case prices of wines from 36 wineries in Upstate New York. Test whether the mean case price is the same for the three Regions (Cayuga, Seneca or Keuka). This is a generalization of the two-sample *t* test, but now there are 3 groups. We could do 3 *t*-tests, but we should then adjust α if we did that because the chance of

not making a Type I error when doing more than one test goes down. For example, if we did 10 *t*-tests each at $\alpha = 0.05$, then P(no Type I error 10 times) $= (1\text{-}0.05)^{10} = .60$, which means that α is really 0.40, not 0.05.

There are several common method for comparing lots of means. One is found by going to **Compare means** (in the red triangle). You'll see four choices. Select **All pairs, Tukey's HSD.** This method adjusts the margin of error to compensate for the fact that you're doing lots of comparisons. (**Each pair, Student's t** would do the 3 standard *t*-tests and not adjust.) What region's means are distinguishable from each other? Use the circles and the chart at the bottom of the output. There is also an overall test that all the means are the same, called the *F*-test, that you can see if you select **Means/ANOVA** from the red triangle and then look at the P-value in the table called Analysis of Variance. The null hypothesis is that all means are equal. The alternative is that any mean is different.

6. **Containers again.** The data set Cupps contains the data from an experiment on heat loss from four containers done by a Williams student. She heated water to 180 degrees F, put it into one of four different types of mug and then measured the heat loss 30 minutes later (Difference in degrees Fahrenheit). She tested each mug 8 times.
   a) What are the null and alternative hypotheses?
   b) Carry out the ANOVA in JMP.
   c) If you rejected the null hypothesis, then at least one container's mean heat loss is different from the rest. Use Tukey's HSD method to see which containers can be distinguished.

7. **Used Cars** Classified ads in the Ithaca Journal offered several used Toyota Corollas for sale. Listed below are the ages of the cars and the advertised prices.

Age (years)	Prices advertised
1	12995, 10950
2	10495
3	10995, 10995
4	6995, 7990
5	8700, 6995
6	5990, 4995
9	3200, 2250, 3995
11	2900, 2995
13	1750

   a) Make a scatterplot for these data.
   b) Do you think a linear model is appropriate?
   c) Find the equation of the regression line.
   d) Check the residuals to see if the conditions for inference are met. (Find the options under the red triangle for **Linear Fit**).

e) Create an approximate 95% confidence interval for the slope of the regression line by finding the standard error of the slope and using 2 standard errors on either side of the estimate.
f) Explain what your confidence interval means.

8. **Body Fat (**The data set body fat contains the data.)
   a) Do these data indicate an association between waist size and body fat index?
   b) Check that the conditions for regression inference are met.
   c) Test an appropriate hypothesis about the association and state your conclusion.
   d) Give a 95% confidence interval for the mean percent body fat found in people with 40" waists. Here's how to do this in JMP. First, at the end of the data set add a row with a 40 " waist. (Don't make up a % body fat, though). Now, instead of **Fit Y by X,** use **Fit Model**. Put in the usual for X and Y. Under the red triangle, you'll find **Save columns** and **Mean Confidence Interval.** Now, in your data set you'll see the upper and lower confidence bounds for the mean weight for every data value (including the one you added at 40").

9. **Body Fat, Again** Do these data indicate an association between *weight* and percent body fat?
   a. Check that the conditions for regression inference are met.
   b. Find a 90% confidence interval for the slope of the line of regression of Body Fat on Weight.
   c. Interpret your interval in context.
   d. Give a 95% prediction interval for the body fat index of the mean of all people who weigh 165 pounds.
   e. Give a 95% prediction interval for the body fat index of an individual who weighs 165 pounds. (This time save the individual confidence interval).
   f. Compare the intervals in d and e. Why is one larger?