IT3030 Final exam 6/26/2018 ID:_____ Name:_____

An OPEN-book test.

1. (20%) A new drug is available for treating a certain disease that requires hospitalization. Taking the number of days for hospitalization as a random variable. The study group of 10 people using this drug have the hospital stay for {9, 14, 11, 3, 9, 10, 8, 4, 23, 14} days. A control group of 8 people did not have this drug treatment have hospital stay for {14, 28, 22, 14, 20, 37, 44, 28} days. (a) Compute the mean and standard deviation of days to stay in the hospital for both groups. (b) Perform a 2-sided t test between the two samples to determine whether the difference of mean days to stay in the hospital is significant or not. (Clearly compute your pooled estimate of the variance from the two groups, the t statistic, the degree of freedom you used for a proper t-distribution, and the p-value for whether or not to reject the null hypothesis that the two means are the same. α=0.05)

>> Study=[9, 14, 11, 3, 9, 10, 8, 4, 23, 14];
>> Control=[14, 28, 22, 14, 20, 37, 44, 28];
>> n1=10;n2=8;
>> x1=mean(Study) = **10.5000**
>> x2=mean(Control) = **25.8750**
>> s1=std(Study) = **5.6814**
>> s2=std(Control) = **10.6427**
>> sp2=((n1-1)*s1^2+(n2-1)*s2^2)/(n1+n2-2) = **67.7109**
>> t=(x1-x2)/sqrt(sp2*(1/n1+1/n2)) = **-3.9391**
>> DF=n1+n2-2 = **16**
>> p = 2*tcdf(t,DF) = **0.0012**

This is smaller than 0.05, we thus **reject** the null hypothesis, and conclude that **the drug indeed changed the days for a hospital stay**.

2. (20%) The score distribution for Biostatistics in 2017 looks like the following:

	60~69	70~79	80~89	90~100
Boys	10	3	5	2
Girls	1	5	0	4

(a) Construct an expected table. Keep number of people to 1 decimal point.

(b) Compute the χ^2 statistic, clearly state the degree of freedom used to choose a proper distribution, and compute the p-value for determining whether the scores between boys and girls are different (assuming α=0.05).

	60~69	70~79	80~89	90~100	
Boys	**7.3**	**5.3**	**3.3**	**4**	20
Girls	**3.7**	**2.7**	**1.7**	**2**	10
	11	8	5	6	

>> O=[10 3 5 2 1 5 0 4];

>> E=[7.3 5.3 3.3 4 3.7 2.7 1.7 2];

>> chi2=sum((O-E).^2./E) = **11.5020**

>> DF=(2-1)*(4-1) = **3**

>> p=1-chi2cdf(chi2,DF) = **0.0093**

This is smaller than 0.05, we thus **reject** the null hypothesis of no difference. That is, the two set of scores are **significantly different**.

3. (30%) A study of a new cancer treatment and how the new treatment extended the days of survival for cancer patients is listed below.

Cancer type	Count	Mean (days)	Standard deviation (days)
Breast	11	1395.9	1239.0
Colon	17	457.4	427.2
Ovary	6	884.3	1098.6
Stomach	13	286.0	346.3

(a) Compute the grand mean of all patients.

(b) Compute S_w^2 and S_B^2, and from them to compute the F statistic. What are the two degrees of freedom to use for this particular F-distribution?

(c) Compute the p-value for the F-test and state your conclusion. Assume α=0.05. [Hint: You should reject the null hypothesis that these groups have "equal" survival days.]

```
>> n1=11;n2=17;n3=6;n4=13;
>> x1=1395.9;x2=457.4;x3=884.3;x4=286.0;
>> s1=1239.0;s2=427.2;s3=1098.6;s4=346.3;
>> n=n1+n2+n3+n4 =    47
>> x=(n1*x1+n2*x2+n3*x3+n4*x4)/n =    684.1383
>> k=4;
>> sw2=((n1-1)*s1^2+(n2-1)*s2^2+(n3-1)*s3^2+(n4-1)*s4^2)/(n-k) =    5.9872e+05
>> sb2=(n1*(x1-x)^2+n2*(x2-x)^2+n3*(x3-x)^2+n4*(x4-x)^2)/(k-1) =    2.9159e+06
>> F=sb2/sw2 =    4.8702
>> DF1=k-1 =    3
>> DF2=n-k =    43
>> p=1-fcdf(F,DF1,DF2) =    0.0053
```

Since this is smaller than 0.05, we **reject the null hypothesis** and conclude that **there exists difference** in these 4 groups in terms of mean survival days after the treatment.

4. (30%) Following problem 3, after you have rejected the null hypothesis. Now you realize there exists difference in some of the paired samples.

(a) What is the level of significance you should use for finding the different pair(s)?

(b) Use the same S_w^2 obtained from Problem 3 for all paired pooled variance. What is the degree of freedom used for these t-tests?

(c) Compute the t statistic and p-values for all 6 pairs. What pair(s) has (or have) significant difference?

New alpha would be 0.05/C(4,2)=**0.0083**

The DF used in these follow-up t-tests would be the same as **DF2=43** above.

Compute the t statistic:

>> t12=(x1-x2)/sqrt(sw2*(1/n1+1/n2)) = **3.1345**

>> t13=(x1-x3)/sqrt(sw2*(1/n1+1/n3)) = **1.3028**

>> t14=(x1-x4)/sqrt(sw2*(1/n1+1/n4)) = **3.5013**

>> t23=(x2-x3)/sqrt(sw2*(1/n2+1/n3)) = **-1.1619**

>> t24=(x2-x4)/sqrt(sw2*(1/n2+1/n4)) = **0.6012**

>> t34=(x3-x4)/sqrt(sw2*(1/n3+1/n4)) = **1.5667**

>> p12=2*(1-tcdf(t12,DF2)) = **0.0031**

>> p13=2*(1-tcdf(t13,DF2)) = **0.1996**

>> p14=2*(1-tcdf(t14,DF2)) = **0.0011**

>> p23=2*tcdf(t23,DF2) = **0.2517**

>> p24=2*(1-tcdf(t24,DF2)) = **0.5508**

>> p34=2*(1-tcdf(t34,DF2)) = **0.1245**

Only p12 and p14 are smaller than adjusted alpha = 0.0083, meaning a significant difference was noticed between **groups 1 vs 2**, as well as **groups 1 vs 4**. No significant difference in the other 4 pairs.

5. (20%) Data from 8 male bears are shown below.

Bear ID	1	2	3	4	5	6	7	8
y Weight	80	344	416	348	262	360	332	34
x2 Age	19	55	81	115	56	51	68	8
x3 Neck	16.0	28.0	31.0	31.5	26.6	27.0	29.0	13.0
x4 Length	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
x5 Chest	26	45	54	49	41	49	44	19

(a) Determine the values for b1~b5 where y=b1+b2*x2+b3*x3+b4*x4+b5*x5 in a multiple regression.

(b) What are the predicted weights for these 8 bears based on your regression function?

```
>> y=[80  344  416  348  262  360  332  34]';
>> x2=[1955    81    115  56    51    68    8]';
>> x3=[16.0    28.0 31.0 31.5 26.6 27.0 29.0 13.0]';
>> x4=[53.0    67.5 72.0 72.0 73.5 68.5 73.0 37.0]';
>> x5=[2645    54    49    41    49    44    19]';
>> A=[ones(size(x2)) x2 x3 x4 x5]
A =
      1.0000    19.0000    16.0000    53.0000    26.0000
      1.0000    55.0000    28.0000    67.5000    45.0000
      1.0000    81.0000    31.0000    72.0000    54.0000
      1.0000   115.0000    31.5000    72.0000    49.0000
      1.0000    56.0000    26.6000    73.5000    41.0000
      1.0000    51.0000    27.0000    68.5000    49.0000
      1.0000    68.0000    29.0000    73.0000    44.0000
      1.0000     8.0000    13.0000    37.0000    19.0000
>> b=regress(y,A)
b = -216.3885   -1.3003   18.9086   -3.2042    7.1244
>>>> yp=b(1)+b(2)*x2+b(3)*x3+b(4)*x4+b(5)*x5
yp =
   76.8556
  345.8512
  418.4702
  348.0923
  270.3559
  357.4372
  323.1084
   35.8293
```