Hands on Assignment #2

1) Use the dataset in the "Class Survey" Excel file, which contains an in-class survey of introductory statistics students. Detailed variable definitions and question wording are included in the file. Answer the following questions, at the 0.1 level of significance.

a. How many males and females are there?

Number of males: n1=92

Number of females: n2=91

b. Create a crosstab of counts of "like cats" by "male or female". Please properly format your table (i.e. no "0" or "1". Use proper labels for "male", "female", "likes cats", "does not like cats").

Gender	Like cats	Don't like cats	Total
Male	45	47	92
Female	50	41	91
Grand Total	95	88	183

c. What proportion of students like cats? What proportion of male students like cats? What proportion of female students like cats?

P(students like cats)= Total number of students who like cats/ Total number of students

=95/183=0.5191

P(male students like cats)= Number of males who like cats/Total number of males

=45/92=0.4891

P(female students like cats)=Number of females who like cats/Total number of females

=50/91=0.5494

d. Create a 99% confidence interval for the proportion of statistics students who like cats and interpret the interval.

Level of significance $\alpha=0.01$; n=183

P= number of students who like cats/total number of students

P=95/183=0.5191

The confidence interval of population proportion:

$$P \pm z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}$$

We can assume that *p* is normally distributed because *np* and *n*(1 − *p*) exceed 10. That is,

np = (183)(0.5191) = 94.99 ; n(1 − p) = (183)(1-0.5191) = 88

$Z_{\alpha/2}$ =2.576

$$P \pm z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}$$

=0.5191+ 0.09514; 0.5191 - 0.09514

The confidence interval of population proportion= 0.61424, 0.42396

We are 99% confident that the true population proportion of students likes cat is between 42.396% and 61.424%.

e. Do an appropriate analysis at the 0.1 level of significance to see whether "liking cats" is a different proportion for males vs. females.

Using α=0.1

1. Claim $\pi 1 \neq \pi 2$
2. H_0: $\pi 1 = \pi 2$
   H_A: $\pi 1 \neq \pi 2$
3. Calculation:
   P1= P(male students like cats)
   P1=x1/n1 = Number of males who like cats/Total number of males
   P1=45/92=0.4891

   P2= P(female students like cats)
   P2=x2/n2==Number of females who like cats/Total number of females
   P2=50/91=0.5494

   Pc=(x1+x2)/(n1+n2)
   Pc=(45+50)/(92+91)= 0.5191

For two -tailed test:

Critical points $Z_{\alpha/2} = \pm 1.645$

Test Statistic:

$Z_{cal} = ((p1 - p2) - (\pi 1 - \pi 2)) \div (\sqrt{\left(pc(1-pc)\left[\frac{1}{n1} + \frac{1}{n2}\right]\right)})$

$\pi 1 - \pi 2 = 0$

$Z_{cal} = ((p1 - p2) - 0) \div (\sqrt{\left(pc(1-pc)\left[\frac{1}{n1} + \frac{1}{n2}\right]\right)})$

$Z_{cal} = -0.8163$

p-value approach:

p-value = 2. P(z<-0.8163) = 2*0.2072 =0.414

Since, p-value> α, we fail to reject H_0.

Summarizing in table

Content	Values
Alpha	0.1
Proportions:	
P1(male)	0.4891
P2(female)	0.5494
Pc(Pooled)	0.5191
2-tail test:	
Z critical value	1.645
Z test statistics	-0.8163
p-value	0.414

Conclusion:

There is no significant difference between males and female proportions who like cats.

f. Create a 90% confidence interval for the difference in proportion between males and females that like cats. Interpret it.

Level of significance α=0.1; n=183

P1= P(male students like cats)

P1=x1/n1 = Number of males who like cats/Total number of males

P1=45/92=0.4891

P2= P(female students like cats)

P2=x2/n2=Number of females who like cats/Total number of females

P2=50/91=0.5494

$Z_{\alpha/2}$ = 1.64485

The confidence interval:

$$(p1 - p2) \pm \text{Z}\alpha/2(\sqrt{\left(\frac{p1(1 - p1)}{\text{n1}} + \frac{p2(1 - p2)}{\text{n2}}\right)}$$

The confidence interval of population proportion= 0.061, -0.1816

-0.1816 <π1-π2< 0.061

We are 90% confident that difference in population proportion between males and females who like cats lie between -18.16% and 6.1%.

g. Compare your answers to (e) and (f) and explain the connection between the two. (Hint: you need to be mentioning whether 0 is in the interval).

Since 90% confidence interval includes 0, we will reject null hypothesis of no difference in proportions, hence unequal population proportions for male and female who like cats. In 95% confidence interval, we are failed to reject null hypothesis of no difference in proportions. Hence, equal population proportions for male and female who like cats.

h. Check the validity of the data in the GPA field and explain how you treat special cases.

GPA should be from 0 to 4. GPA is not mentioned for 3 males and 1 female. We will not include these 3 males and one female in our calculation. There is one special case which shows GPA as 25.0. We will remove this outlier before performing any hypothesis test.

i. What's the average GPA? What's the average GPA for males? What's the average GPA for females?

We are excluding outlier 25 as GPA can't be greater than 4.

Table1:

Gender	Number of students	Average GPA	Standard deviation
Male	89	3.0924	0.4373

Female	90	3.2122	0.3836
Grand total	179	3.1526	0.4144

j. Create a 95% confidence interval for the GPA of statistics students and interpret the interval.

Level of significance α=0.05

Number of students n=179

Degree of freedom =n-1=178

Confidence interval for mean μ (with unknown σ)

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

$t_{\alpha/2}$=T.INV.2T(0.05, 178)=1.9734

$\bar{x}$=Average GPA of students=3.1526 (From Table1)

S= sample standard deviation=0.4144 (From Table1)

Confidence interval=3.1526 ± 0.06112

3.21372, 3.09148

We are 95% confident that GPA of statistics students will fall within 3.09148 and 3.21372.

k. Perform an appropriate analysis (F test) to see if there is a difference in variability in GPA for males vs. females at the 0.1 level of significance.

F- test for two-tailed test:

1. Claim $\sigma_1^2 \neq \sigma_2^2$

2. H_0: $\sigma_1^2 = \sigma_2^2$
   H_A: $\sigma_1^2 \neq \sigma_2^2$

3. Degrees of freedom are:
Numerator: df1 = n1 – 1 =89-1 =88
Denominator: df2 = n2 – 1=90-1=89

n1=number of males who got GPA

n2=number of females who got GPA

4. Fcalc = s_1^2/s_2^2
   S1= Sample standard deviation of GPA for males=0.4373
   S2= Sample standard deviation of GPA for females=0.3836
   Fcalc=0.1912/0.1472=1.2996
5. Fcritical=upper tail value=1.3146
   <u>P value approach:</u>
   p-value =2*F.DIST.RT(1.2996, 88,89)=2*0.1097=0.2194
   p-value>level of significance. We fail to reject H0.

F-Test Two-Sample for Variances		
	Male	*Female*
Mean	3.0924	3.2122
Variance	0.1912	0.1472
Observations	89	90
df	88	89
F	1.2996	
P(F<=f) one-tail	0.1097	
P(F<=f) two-tail	0.2194	
F Critical upper-tail	1.3146	

Since Fcalc<Fcritical, we fail to reject H0.

Conclusion: There is no significant differences in variances of GPA for males and females at the level of significance 0.1.

l. Perform an appropriate analysis with an appropriate t test to see if there a difference in GPA for males vs. females, at the 0.1 level of significance. (Hint: Use the results of the F test to know which t test to use).

From F-test, we conclude that there is no significant differences in variances of GPA for male and female. We will conduct t-test assuming equal variances of GPA for male and female.

Hypothesis test: Check if there is significant difference in population mean of GPA for male vs female.

1. Claim, $\mu_1 - \mu_2 \neq 0$
2. H0: $\mu_1 - \mu_2 = 0$
   HA: $\mu_1 - \mu_2 \neq 0$
3. Degrees of freedom : df=n1+n2-2
   n1=number of males=89
   n2=number of females=90

df=89+90-2=177

Pooled variance=Sp

$$\text{Sp} = \frac{s_1^2(n_1-1)+S_2^2(n_2-1)}{n_1+n_2-2}$$

$$\text{tcalc} = \left(\frac{\bar{x}_1-\bar{x}_2-(\mu_1-\mu_2)}{\sqrt{s_p^2/n_1 + s_p^2/n_2}} \right)$$

tcritical= =T.INV.2T(0.1,177)= 1.6535

t-Test: Two-Sample Assuming Equal Variances		
	Male	*Female*
Mean	3.0924	3.2122
Variance	0.1912	0.1472
Observations	89	90
Pooled Variance	0.1691	
Hypothesized Mean Difference	0	
df	177	
t Stat	-1.9500	
P(T<=t) one-tail	0.0264	
t Critical one-tail	1.2864	
P(T<=t) two-tail	0.0528	

t Critical two-tail	1.6535	

tcalc = -1.95

tcritical for two tail=1.6535

t-critical value for lower tail (in two-tail test)=-1.6535

tcalc will lie in left side rejection region. Hence, we reject H0.

Conclusion: There is significant difference in GPA in population for male and female.

m. Come up with and perform your own hypothesis test on the data, using the field "carAge". Make sure that you do a validity check on this field before beginning, as you did with GPA. Please make your write up as formal and clear as possible.

For CarAGe: We have removed outlier 95 and 99 before performing calculations.

Hypothesis test:

Check if there is difference in population mean of CarAge for male and female at level of significance 0.1

We will perform t-test for unknown variances assumed equal:

1. Claim, $\mu_1 - \mu_2 \neq 0$
2. H0: $\mu_1 - \mu_2 = 0$
   HA: $\mu_1 - \mu_2 \neq 0$
3. Degrees of freedom : df=n1+n2-2
   n1=number of males=88
   n2=number of females=91

df=88+91-2=177

Pooled variance=Sp

$$\text{Sp} = \frac{s_1^2(n_1-1)+S_2^2(n_2-1)}{n_1+n_2-2}$$

$$\text{tcalc} = \left(\frac{\bar{x}_1-\bar{x}_2-(\mu_1-\mu_2)}{\sqrt{s_p^2/n_1 + s_p^2/n_2}} \right)$$

tcritical= =T.INV.2T(0.1,177)= 1.6535

t-Test: Two-Sample Assuming Equal Variances		
	Male	*Female*
Mean	5.040	3.9945
Variance	12.829	12.9250
Observations	88	91
Pooled Variance	12.878	
Hypothesized Mean Difference	0	
df	177	
t Stat	1.948	
P(T<=t) one-tail	0.026	
t Critical one-tail	1.286	
P(T<=t) two-tail	0.053	
t Critical two-tail	1.654	

tcalc=1.948

tcritical for two tail =1.654

Since tcalc>tcritical, we reject H0.

Conclusion: There is significant difference in population mean of CarAge for male and female.

2) This is an experiment to illustrate the Central Limit Theorem.

a. State the Central Limit Theorem and explain what it means in your own words.

The Central limit theorem allow us to estimate the sampling distribution of sample mean $\bar{X}$ when we don't have idea about the distribution of population. The Central limit theorem states that if a random sample of size n is drawn from a population with mean μ and standard deviation σ, the distribution of the sample mean $\bar{X}$ approaches a normal distribution with mean μ and standard deviation $\sigma_{\bar{x}}$ = σ/sq root(n) as sample size increases.

We may have population distribution as normal, triangular, uniform, skewed etc. By the central limit theorem, if sample size is large enough, normal distribution will be the shape of sample mean. Much smaller sample size will be sufficed if population is symmetric.

b. Use =RANDBETWEEN(1, 100) to create 25 samples of size n=9 by choosing two-digit random numbers between 1 and 100.

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9
100	60	99	56	77	12	29	26	74
52	18	67	2	51	90	59	46	82
72	56	5	66	90	56	3	82	20
1	60	4	97	47	12	96	10	44
93	73	96	2	41	90	83	87	33
93	93	18	49	79	70	30	98	100
65	72	16	55	95	38	1	49	25
1	82	9	43	86	56	88	36	17
54	81	12	3	12	16	42	72	59
21	7	68	18	98	90	41	80	40
81	17	49	11	26	71	10	100	72
48	29	31	27	100	26	64	28	12
84	28	68	48	62	25	78	20	11
38	60	99	28	28	47	56	73	74
81	91	74	53	26	33	24	94	19
83	48	51	84	52	71	99	42	50
26	45	13	60	56	88	39	77	53
100	66	23	19	59	82	53	84	3
46	76	49	97	76	70	76	36	50
52	7	98	54	1	82	2	60	17
33	62	6	83	16	89	10	74	5
42	10	95	67	94	78	14	89	12
97	74	22	40	77	18	56	44	10
71	76	6	80	83	63	74	54	35
91	12	40	26	33	57	39	53	36

c. What distribution are you using when you do this? What are the mean and SD of this distribution?

We are using Discrete Uniform distribution.

$X \sim_{\text{Discrete}} U(a=1, b=100)$

Parameters:

Lower Limit a=1, Upper Limit b=100

Mean μ = (a+b)/2;

Theoretical value: Mean=50.5

Observed Value: Mean =51.742

Standard deviation=SQRT(((b-a+1)^2-1)/12); Standard deviation=28.866

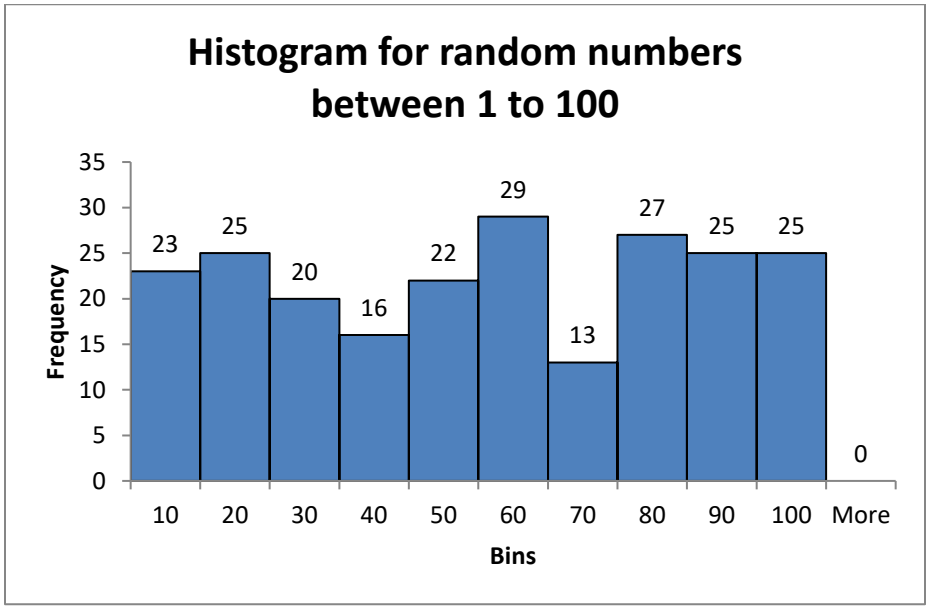
d. For each sample, calculate the mean.

Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample Mean
100	60	99	56	77	12	29	26	74	59.222
52	18	67	2	51	90	59	46	82	51.889
72	56	5	66	90	56	3	82	20	50.000
1	60	4	97	47	12	96	10	44	41.222
93	73	96	2	41	90	83	87	33	66.444
93	93	18	49	79	70	30	98	100	70.000
65	72	16	55	95	38	1	49	25	46.222
1	82	9	43	86	56	88	36	17	46.444
54	81	12	3	12	16	42	72	59	39.000
21	7	68	18	98	90	41	80	40	51.444
81	17	49	11	26	71	10	100	72	48.556
48	29	31	27	100	26	64	28	12	40.556
84	28	68	48	62	25	78	20	11	47.111
38	60	99	28	28	47	56	73	74	55.889
81	91	74	53	26	33	24	94	19	55.000
83	48	51	84	52	71	99	42	50	64.444
26	45	13	60	56	88	39	77	53	50.778
100	66	23	19	59	82	53	84	3	54.333
46	76	49	97	76	70	76	36	50	64.000
52	7	98	54	1	82	2	60	17	41.444
33	62	6	83	16	89	10	74	5	42.000
42	10	95	67	94	78	14	89	12	55.667
97	74	22	40	77	18	56	44	10	48.667
71	76	6	80	83	63	74	54	35	60.222
91	12	40	26	33	57	39	53	36	43.000

e. Make a histogram of the 225 individual x-values using bins 10 units wide. Describe the shape of the histogram.

Bin	*Frequency*
10	23
20	25
30	20
40	16
50	22
60	29
70	13
80	27
90	25
100	25
More	0

Histogram:



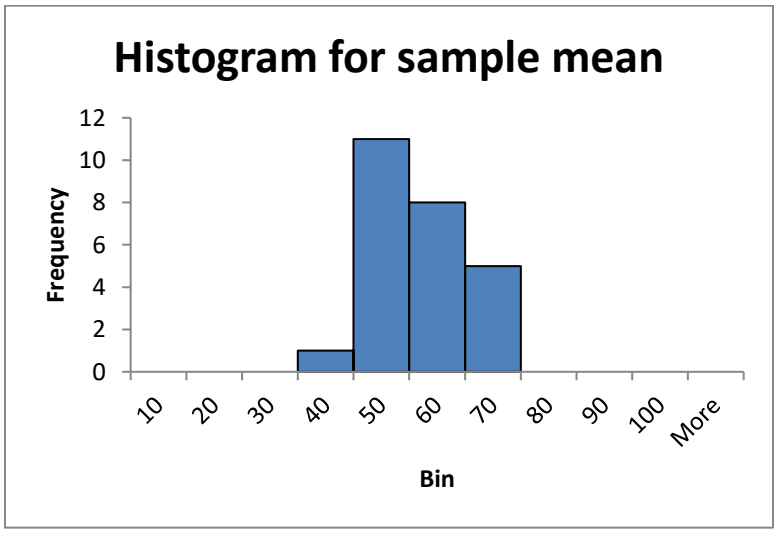
The shape of the distribution does not look bell-shaped. It looks uniform distribution with mean 51.742 and standard deviation 28.866.

f. Make a histogram of your 25 sample means using bins 10 units wide.

Sample Mean	Bin	Frequency
59.222	10	0
51.889	20	0

50.000	30	0
41.222	40	1
66.444	50	11
70.000	60	8
46.222	70	5
46.444	80	0
39.000	90	0
51.444	100	0
48.556	More	0
40.556		
47.111		
55.889		
55.000		
64.444		
50.778		
54.333		
64.000		
41.444		
42.000		
55.667		
48.667		
60.222		
43.000		

Histogram:



g. Discuss the histogram shape. Does the Central Limit Theorem seem to be working?

The shape of the distribution does not look quite normal as sample size was 9 which is small. If sample size will be 30 or more, the shape will tend to obtain normal distribution. The mean of the distribution is 51.7422 with standard deviation 8.7037.

If sample size would be 30 or more, Central limit theorem will work and shape of histogram will be normal.

h. Find the mean of your 25sample means. What value should it be, according to the central limit theorem? Was it what you would expect?

<u>The average of Sample Mean:</u>

Observed value from the data: $\mu_{\bar{x}} = 51.7422$

According to Central limit theorem,

Expected value (Theoretical value): E(X)=50.5

Expectation: The observed value of the average of sample mean (51.742) should be close to theoretical value (50.5), which is true in this case.

i. Find the standard deviation of your 25 sample means. What value should it be, according to the central limit theorem? Was it what you would expect?

Theoretical Parameter:

Lower Limit a=1, Upper Limit b=100

The standard deviation of the sample means

Standard deviation $\sigma_{\bar{x}}$ =8.7037(Observed)

According to central limit theorem, Standard deviation is:

$\sigma_{\bar{x}}$ =9.622 (Theoretical)

Expectation: The observed value of standard deviation should be close to theoretical value. In this case, both values are closer.