Vy Pham
ID: 933-313-965

16 April 2019
To: CBEE 213 Class From:
Professor Koretsky
Subject: Homework #3

Below is the third homework assignment of the quarter. This assignment is due 24 April 2019 by 8 AM on Gradescope. You will submit one file with your computer work and another file where you report your answers. Please conform to the format described on the class web site. http://classes.engr.oregonstate.edu/cbee/spring2019/cbee213-010/HWFormat.pdf If you have any questions, feel free to see me or one of the other instructors during office hours or by appointment.

Data for the problems is available below the link for this HW

1. **An article in the *Transactions of the American Fisheries Society* reports the results of a study to investigate the mercury contamination in largemouth bass. A sample of fish was selected from 53 Florida lakes and mercury concentration in the muscle tissue was measured (ppm). The mercury concentration values are on the attached .xls file and reported below.**

   A. **To 95% confidence, what do you estimate the mean of Hg in this population to be?**
      To 95% confidence, the mean of Hg in this population is 0.525±0.096 ppm. In another word, the mean of Hg is between 0.429 and 0.621 ppm.

   B. **To 95% confidence, what do you estimate the standard deviation of Hg in this population to be?**
      To 95% confidence, the standard deviation of Hg in this population is between 0.293 and 0.431ppm

      ***Below is my Excel table**

		Function used
mean	0.525	AVERAGE(A1:A53)
stdev	0.349	STDEV(A1:A53)
number of sample	53.000	COUNT(A1:A53)
alpha	0.025	
t-values	2.007	T.INV(1-H12,H17)
mean CI_high	0.621	H9+H13*H16
mean CI_low	0.429	H9-H13*H16
standard error	0.048	H10/SQRT(H11)
DOF	52	N-1
upper_chi	33.968	CHIINV(1-alpha/2,n-1)
lower_chi	73.81	CHIINV(alpha/2,n-1)

Vy Pham
ID: 933-313-965

stdev CI_upper	0.431	H10*SQRT(H17/H18)
stdev CI_lower	0.293	H10*SQRT(H17/H19)

I used the =MEAN(data) and =STDEV(data) function to calculate the mean and the standard deviation of the data respectively. Since we are looking for 95% confidence interval, the alpha/2 value is 0.025 (alpha=0.05). The t values is found by using the =T.INV(1-α/2, DOF) where DOF is found by difference of the number of data (n) and 1, DOF=n-1. The mean 95% confidence interval can be calculated by using the $CI_{high/low} = mean \pm \frac{\sigma}{\sqrt{n}}$. For upper chi-value, I use function =CHIINV(1- α/2, DOF). For the lower chi-value, I use =CHIINV(α/2, DOF). To calculate the upper and lower standard deviation confidence intervals, I use the following function: $CI_{high} = \sigma\sqrt{\frac{n-1}{\chi^2_{high}}}$ and $CI_{low} = \sigma\sqrt{\frac{n-1}{\chi^2_{low}}}$

C. **Comment on the results.**

1.230	0.500	0.410	0.870	0.860
1.330	0.490	0.730	0.560	0.520
0.040	1.160	0.590	0.170	0.650
0.044	0.050	0.340	0.180	0.270
1.200	0.150	0.340	0.190	0.940
0.270	0.190	0.840	0.040	0.400
0.490	0.770	0.500	0.490	0.430
0.190	1.080	0.340	1.100	0.250
0.830	0.980	0.280	0.160	0.270
0.810	0.630	0.340	0.100	
0.710	0.560	0.750	0.210	

For every large number of fish samples collected, we are 95% confident that the population mean for the measured mercury concentration is between 0.429 and 0.621ppm. We are 95% confident that the standard deviation will fall within 0.293 and 0.431 ppm. Since there are high chi-values, meaning that there is a high variability in the data. The mean and the standard deviation were within the captured 95% confidence intervals respectively. Moreover, I haven't given any other related data set so I can just give what I have observed from this specific data.

2. **The class coin flip data are available in the file Studio2_Data.xlsx located below this HW. For each studio section, calculate a 95% confidence interval. How many intervals capture the true population mean?**

From what I have analyzed, all studio sections capture the true population mean

*Below is the table describing explicitly the statistic (mean, standard deviation, and confidence interval) for each section

Vy Pham
ID: 933-313-965

	BEXL 102	BEXL 103	BEXL 102	BEXL 103	BEXL 102	BEXL 103	BEXL 102	BEXL 103	function used
	W 8-10	W 8-10	W 12-2	W 12-2	W 2-4	W 2-4	W 4-6	W 4-6	
mean	2.42	2.08	2.05	2.05	2.05	2.05	2.01	2.06	MEAN()
stdev	1.28	0.87	0.73	0.92	1.28	0.75	1.50	0.84	STDEV()
alpha/2	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025	
t-values	2.01	1.99	1.98	1.98	1.98	1.98	1.99	1.98	T.INV(1-alpha/2, n-1)
CI_high	2.78	2.26	2.18	2.23	2.29	2.19	2.33	2.23	
CI_low	2.06	1.90	1.92	1.88	1.80	1.90	1.70	1.89	
true mean	2.10								MEAN()
True ?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	

*<u>Description for Excel table</u>

To determine if the population capture the true population mean, I need to find the 95% confidence interval for the distribution of the mean.

First, I calculate the number of total measurements (n) for each studio section by taking the sum (using =SUM(data) function) of frequencies (occurrences) for each section. Then, for each studio section, I take the product of each frequency with the corresponding number of heads, and take the sum of each product to generate the total heads. The mean is calculated by taking the sum of heads divided by the amount of total measurements.

To find the standard deviation, for each section, I used the following function $\sigma = \sqrt{\frac{\sum f*(x_i-\mu)^2}{\sum f}}$

Since we are looking for the confidence interval for distribution of mean, alpha/2= 0.025. After having the mean and standard deviation for each studio section, t-values can be found by using the =T.INV(1-α/2, n-1), using 1-α /2 since we want to look to the left of the distribution graph. The 95% confidence interval for the distribution of the mean can be calculated by using the $CI_{high/low} = mean \pm \frac{\sigma}{\sqrt{n}}$ equation. The true mean is generated from the studio means. Check the true mean value with the confidence interval for each section, if the mean is within the interval, say yes that the intervals capture the true population mean and vice versa.

3. **Go to the file <u>http://jimi.cbee.oregonstate.edu/statistics/coin.php?file=coinskew</u> which generates a sample of 100 samples of the experiment you did in Studio 2 where you recorded the number of heads in 4 flips. However, you suspect the coin might be weighted to favor one side. To 95% confidence, what is the population mean? <u>Be sure to choose the Section you are registered in.</u>**

*Below is the table showing the Studio 2 data (W4 Catherine)

Vy Pham
ID: 933-313-965

		Function used		
# coin flip/trial	4			
# trials /sample	100			
#sample	1			
total flips	400			
average	3.11			
stdev	0.92			
alpha/2	0.025			
t-vlues	1.98	T.INV(1-alpha/2,100-1)		
standard error	0.046	standard error=stdev/sqrt(n)		
CI_high	3.201			
CI_low	3.019			

To find t-values, I used function T.INV(1-α/2, (#trials/sample) -1). I used the sample size, trial/sample=100 as my n value. The 95% confidence interval for the distribution of the mean can be calculated by using the $CI_{high/low} = mean \pm \frac{\sigma}{\sqrt{n}}$ equation.

The population mean is 3.11±0.0913, in another way, the population mean is between 3.019 and 3.201. Comparing to my studio data (BEXL 103 4-6) on problem 2, this confidence intervals and the average are skewed toward the left. This implies this coin is weighed toward tail.

4. **For the 1970 draft data we worked with in HW 1, perform a paired t-distribution analysis as follows:**
   A. **Match data from the first 183 days of the year to the last 183 days of the year**
   B. **Take the difference (e,g,, day 184 – day 1)**

DayNo	Draft No	DayNo	Draft No	Difference
1	305	184	350	45
2	159	185	115	44
3	251	186	279	28
4	215	187	188	27
5	101	188	327	226
6	224	189	50	174
7	306	190	13	293
8	199	191	277	78
9	194	192	284	90
10	325	193	248	77
11	329	194	15	314

DayNo	Draft No	DayNo	Draft No	Difference
93	271	276	125	146
94	83	277	244	161
95	81	278	202	121
96	269	279	24	245
97	253	280	87	166
98	147	281	234	87
99	312	282	283	29
100	219	283	342	123
101	218	284	220	2
102	14	285	237	223
103	346	286	72	274

Vy Pham
ID: 933-313-965

12	221	195	42	179
13	318	196	331	13
14	238	197	322	84
15	17	198	120	103
16	121	199	98	23
17	235	200	190	45
18	140	201	227	87
19	58	202	187	129
20	280	203	27	253
21	186	204	153	33
22	337	205	172	165
23	118	206	23	95
24	59	207	67	8
25	52	208	303	251
26	92	209	289	197
27	355	210	88	267
28	77	211	270	193
29	349	212	287	62
30	164	213	193	29
31	221	214	111	110
32	86	215	45	41
33	144	216	261	117
34	297	217	145	152
35	210	218	54	156
36	214	219	114	100
37	347	220	168	179
38	91	221	48	43
39	181	222	106	75
40	338	223	21	317
41	216	224	324	108
42	150	225	142	8
43	68	226	307	239
44	152	227	198	46
45	4	228	102	98
46	89	229	44	45
47	212	230	154	58
48	189	231	141	48
49	292	232	311	19
50	25	233	344	319
51	302	234	291	11
52	363	235	339	24
53	290	236	116	174
54	57	237	36	21
55	236	238	286	50
56	179	239	245	66
57	365	240	352	13
58	205	241	167	38

104	124	287	138	14
105	231	288	294	63
106	273	289	171	102
107	148	290	254	106
108	260	291	288	28
109	90	292	5	85
110	336	293	241	95
111	345	294	192	153
112	62	295	243	181
113	316	296	117	199
114	252	297	201	51
115	2	298	196	194
116	351	299	176	175
117	340	300	7	333
118	74	301	264	190
119	262	302	94	168
120	191	303	229	38
121	208	304	38	170
122	330	305	79	251
123	298	306	19	279
124	40	307	34	6
125	276	308	348	72
126	364	309	266	98
127	155	310	310	155
128	35	311	76	41
129	321	312	51	270
130	197	313	97	100
131	65	314	80	15
132	37	315	282	245
133	133	316	46	87
134	295	317	66	229
135	178	318	126	52
136	130	319	127	3
137	55	320	131	76
138	112	321	107	5
139	278	322	143	135
140	75	323	146	71
141	183	324	203	20
142	250	325	185	65
143	326	326	156	170
144	319	327	9	310
145	31	328	182	151
146	361	329	230	131
147	357	330	132	225
148	296	331	309	13
149	308	332	47	261
150	226	333	281	55

Vy Pham
ID: 933-313-965

59	299	242	61	238
60	285	243	333	48
61	108	244	11	97
62	29	245	225	196
63	267	246	161	106
64	275	247	49	226
65	293	248	232	61
66	139	249	82	57
67	122	250	6	116
68	213	251	8	205
69	317	252	184	133
70	323	253	263	60
71	136	254	71	65
72	300	255	158	142
73	259	256	242	17
74	354	257	175	179
75	169	258	1	168
76	166	259	113	53
77	33	260	207	174
78	332	261	255	77
79	200	262	246	46
80	239	263	177	62
81	334	264	63	271
82	265	265	204	61
83	256	266	160	96
84	258	267	119	139
85	343	268	195	148
86	170	269	149	21
87	268	270	18	250
88	223	271	233	10
89	362	272	257	105
90	217	273	151	66
91	30	274	315	285
92	32	275	359	327

151	103	334	99	4
152	313	335	174	139
153	249	336	129	120
154	228	337	328	100
155	301	338	157	144
156	20	339	165	145
157	28	340	56	28
158	110	341	10	100
159	85	342	12	73
160	366	343	105	261
161	335	344	43	292
162	206	345	41	165
163	134	346	39	95
164	272	347	314	42
165	69	348	163	94
166	356	349	26	330
167	180	350	320	140
168	274	351	96	178
169	73	352	304	231
170	341	353	128	213
171	104	354	240	136
172	360	355	135	225
173	60	356	70	10
174	247	357	53	194
175	109	358	162	53
176	358	359	95	263
177	137	360	84	53
178	22	361	173	151
179	64	362	78	14
180	222	363	123	99
181	353	364	16	337
182	209	365	3	206
183	93	366	100	7

C. **Find the confidence interval of the difference.**

*****Here is the Exel data table

		function used
Mean	124.3005	MEAN
Stdev	88.98192	STDEV
alpha/2	0.025	
n	183	
Standard error	6.577733	stdev/sqrt(n)
t-values	1.973084	T.INV(1-alpha/2,n-1)

Vy Pham
ID: 933-313-965

CI_high	137.279	
CI_low	111.3221	

I distributed the first half and the second half equally, and there are a total of 183 values for the difference of the first half and second half. Then I take the absolute value of the difference of the first half and the second half (e.g, =ABS(day 184- day1)). I found the mean and standard deviation of the difference by using =MEAN(data) and =STDEV(data) functions, respectively. Then I use equation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ to find the standard error. In order to find t-value, I use =T.INV(1-α/2, n-1). After getting the t-values, I look for the confidence interval by using the following equation $CI_{high/low} = mean \pm \frac{\sigma}{\sqrt{n}}$.

D. **Does it suggest statistical evidence that the draft was not fair? Explain**

The statistical evidence suggest that the draft was not fair because the difference between the first half and the second half was exactly 0 throughout. Also, the value of standard error of the difference (between first half and second half) was high (approximately 6.58), and the higher the standard error is, the more uncertainty the population. Moreover, many of the differences does not fall between the 95% confidence interval (from 111.3 and 137.3). Therefore, it is true to say that the draft was not fair.

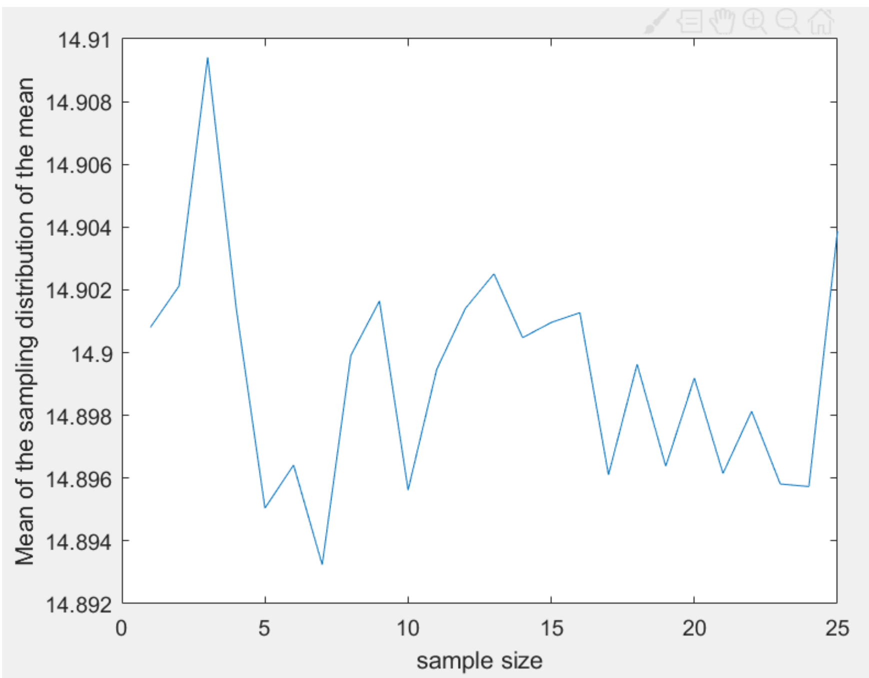
5. **Modify your MATLAB code that you developed in Studio 3 to collect data from samples ranging from 1 measurement to 25 measurements (you did the case of 4 measurements in the studio).**

*****Below is my MatLab code

```
%preamble
clc; close all; clear all;

%opens rows 1:80 of the raw data on the excel file
sampleset=xlsread("Studio_3_Data.xlsx","A1:A80");

%Input number of measurement and sample
n_sample=10000;
%calculates the mean and standard deviation of the raw sample set
x=mean(sampleset);
s=std(sampleset);

%creates random number normally distributed with mean "x" and st. deviation
%"s" calculated above
for i=1:25
data=x+s*randn(i,n_sample);

%calculates the mean and standard deviation of each of each sample that
%contains 4 measurements
avg=mean(data);
stdev=std(data);

%calc mean of sampling dist
```

```
mean_A(i)=mean(avg);
stdev_B(i)=std(avg);
end

%plot figure 1
figure(1)
plot(1:25,mean_A)
xlabel('sample size')
ylabel('Mean of the sampling distribution of the mean')

%plot figure 2
figure(2)
plot(1:25,stdev_B)
xlabel('sample size')
ylabel('Standard deviation of the sampling distribution of the mean')
```

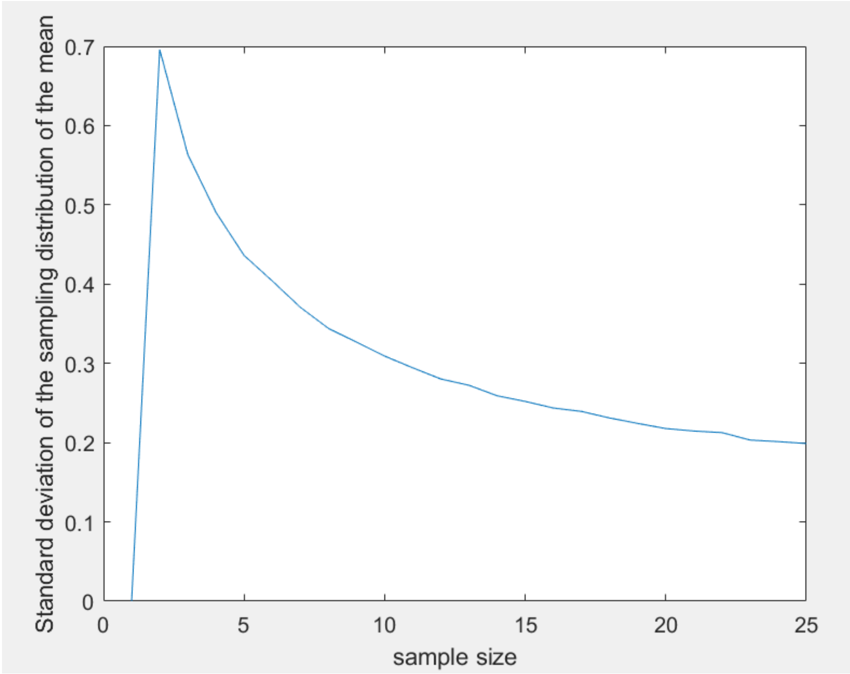
A. **Repeating the sampling 10,000 times come up with an estimate of the mean of the sampling distribution of the mean as a function of sample size for sample sizes ranging from 1 to 25. Plot the result.**



Sampling distribution of the mean

B. **Repeat for the standard deviation of the sampling distribution of the mean.**

Vy Pham
ID: 933-313-965



Sampling distribution of the variance