

ETC1000 Business and Economic Statistics
Homework Week 4
(to be completed before your Computer Lab class in Week 5)

This homework is designed to help you:

- Use Excel to obtain confidence intervals for the mean, the proportion and coefficient of correlation.
- Interpret confidence intervals for the mean, the proportion and coefficient of correlation.
- Understand the theory behind construction of confidence intervals, and appreciate their importance.

1. Confidence Interval for the Mean

The sample mean, $\bar{X}$, is a sample statistic used to estimate the population parameter, μ . A different sample will produce a different $\bar{X}$, so it is important to understand how much $\bar{X}$ might vary from sample to sample, and thus how confident we are in our estimate of μ .

A confidence interval for the mean is a range of values within which we are pretty sure μ lies – there is only one true population mean, μ . Let us go through an example to show how to construct and interpret confidence interval estimates for the mean.

Some fast-food chains offer a lower-priced combination meal in an effort to attract budget-conscious customers. One chain test marketed a burger, fries and drink combination for \$5.95. The weekly sales volume for these meals was impressive. Suppose the chain wants to estimate the average amount its customers spent on a meal at their restaurant while this combination offer was in effect. An analyst gathers data on total amount spent by 28 randomly selected customers, which are provided in *Sheet 1* of the Excel file “Homework 4.xlsx”.

1.1. Confidence Interval for the Mean

We can use some Excel formulas to construct a 95% confidence interval for the mean amount spent by customers at this fast-food chain. First note the mathematical formulas we need to construct the confidence interval:

$$\bar{X} - t_{0.025, n-1} S / \sqrt{n} \quad \text{and} \quad \bar{X} + t_{0.025, n-1} S / \sqrt{n}$$

There are 4 values that we need to calculate in these formulas:

$\bar{X}$ is the sample mean, the average amount spent among customers in the sample. We can calculate the sample mean $\bar{X}$ using Excel’s =AVERAGE formula. The sample mean for this sample is \$6.87.

n is the sample size, in this case 28.

$t_{0.025, n-1}$ is the value from the Students t distribution, with 0.025 in each of the tails, and n-1 degrees of freedom. In Excel we calculate this using the formula `=T.INV.2T(0.05,27)` which `=2.05`. Note that while we write $t_{0.025, n-1}$, the probability that needs to be entered into Excel is 0.05 – Excel assumes that this 0.05 is split equally across both tails of the t distribution. Since this can be easily confused, it might be helpful to remember that the value from the standard normal distribution would be 1.96, so our value we obtain in Excel for $t_{0.025, n-1}$ should be fairly close to 1.96 (the larger the sample, the closer it should be to 1.96).

s is the sample standard deviation. We calculate this in Excel using the function `=STDEV.S(B2:B29)` which works out to be `=0.8699`.

Let's put all these values together:

POISSON....						
=F1+F3*F4						
	A	B	C	D	E	F
1	Customer ID	Amount Spent \$	Mean		Sample Mean (X-bar)	\$ 6.87
2	1	6.25	1		Sample Standard Deviation (s)	0.869917
3	2	7.8	1		TINV(0.05,27)	2.051831
4	3	5.5	1		s/sqrt(n)	0.164399
5	4	6.75	1		Lower 95% CI	\$ 6.54
6	5	8.25	1		Upper 95% CI	=F1+F3*F4
7	6	6.4	1			

The values we obtain for the lower and upper 95% confidence intervals give us our 95% confidence interval for μ : the two values tell us that we can be 95% sure that the true average amount spent on a meal is between \$6.54 and \$7.21.

Using the COUNTIFS function

You have already seen the regular `=COUNTIF` function in Excel. An extension to this function is `=COUNTIFS`. It works the same way as `=COUNTIF`, counting the number of cells that satisfy the criteria you specify, however allows for multiple criteria. In particular, we can use the `=COUNTIFS` function to check whether a value falls inside a range specified by two cut-off values.

For example, we could check how many of the 28 customers in our sample spent an amount within this confidence interval range – that is, how many spent between \$6.54 and \$7.21. Manually scanning down the list would be time-consuming and potentially erroneous, so a function like `=COUNTIFS` can be very useful. Here's how we would do it:

POISSON....									
=COUNTIFS(B2:B29,">"&F5,B2:B29,"<"&F6)									
	A	B	C	D	E	F	G	H	I
1	Customer ID	Amount Spent \$	Mean		Sample Mean (X-bar)	\$ 6.87			
2	1	6.25	1		Sample Standard Deviation (s)	0.869917			
3	2	7.8	1		T.INV.2T(0.05,27)	2.051831			
4	3	5.5	1		s/sqrt(n)	0.164399			
5	4	6.75	1		Lower 95% CI	\$ 6.54			
6	5	8.25	1		Upper 95% CI	\$ 7.21			
7	6	6.4	1		COUNTIFS	=COUNTIFS(B2:B29,">"&F5,B2:B29,"<"&F6)			
8	7	5.5	1						

Here we are asking Excel to count the number of cells (customers) whose total spend was above \$6.54 AND below \$7.21: there are 6 of them. Notice that to refer to a value in a cell in our spreadsheet (e.g. the lower value in cell F5), we need to use ">"&F5. If we just wanted to use the value 6.54 directly in our formula, we would use ">6.54". However referencing the cell in the spreadsheet ensures we don't suffer from any rounding discrepancies.

1.2. Confidence Interval for the Mean Using the Regression Data Analysis Tool

We could have used Excel's in-built *Regression* Data Analysis tool to obtain our confidence interval. Let's do this now.

Step 1: Add a column of "1"s called "Mean" alongside the data:

	A	B	C
1	Customer ID	Amount Spent \$	Mean
2	1	6.25	1
3	2	7.8	1
4	3	5.5	1
24	23	7.95	1
25	24	6.7	1
26	25	5.5	1
27	26	7.7	1
28	27	6.95	1
29	28	7.25	1

Step 2: On the *Data* tab, go to *Data Analysis*, and select *Regression*. Complete the dialog box as follows:

Regression ? X

Input

Input Y Range: ↑

Input X Range: ↑

☒ Labels ☒ Constant is Zero

☒ Confidence Level: %

Output options

☒ Output Range: ↑

☐ New Worksheet Ply:

☐ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK Cancel Help

N.B. Be sure to tick the *Constant is Zero* checkbox and, for this example, check the *Confidence Level* is 95% (the default).

You should obtain the following output:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.992364883					
R Square	0.984788061					
Adjusted R Square	0.947751024					
Standard Error	0.86991721					
Observations	28					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	1322.750089	1322.75	1747.922	2.24781E-25	
Residual	27	20.43241071	0.756756			
Total	28	1343.1825				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
Mean	6.873214286	0.1643989	41.80815	4.41E-26	6.535895606	7.210532965

The mean for the sample data is given in the "Coefficients" column – we see that the mean average amount spent by customers in our sample is \$6.87, as calculated with Excel's =AVERAGE formula. This is $\bar{X}$, our point estimate for the true population mean, μ .

The "Lower 95%" and "Upper 95%" columns give us our 95% confidence interval for μ : the two values tell us that we can be 95% sure that the true average amount spent on a meal is between \$6.54 and \$7.21.

What if we wanted to use a different confidence level, say 90% instead of 95%? We can get Excel to do this by changing the *Confidence Level* in the regression dialog box to 90. Try this yourself – the mean should remain the same (\$6.87), but the interval should become narrower around this mean (\$6.59; \$7.15): we can be 90% sure that the true average amount spent by customers is between \$6.59 and \$7.15.

Now try a larger confidence level. Can you predict whether the interval will be wider or narrower than the default 95% case? Why do you think the interval needs to be wider or narrower? And what will happen if we want to construct a 100% confidence interval? In other words, what is the cost if we want the confidence interval constructed to contain the true mean with certainty?

2. Confidence Interval for the Proportion

Now suppose instead of being interested in the mean amount spent at the chain by customers, we are interested in the proportion that opted to purchase the special meal.

Analogous to the sample mean, the sample proportion, p , is a sample statistic used to estimate the population parameter, in this case, π . A different sample will produce a different value for p , so it is important to understand how much p might vary from sample to sample, and thus how confident we are in our estimate of π .

Sheet 2 provides the same customer data, this time indicating whether or not the customer bought the special meal (a value of "1" means they did buy the special meal; a value of "0" means they did not).

Calculate the number of customers in the sample who opted for the special meal using Excel's =COUNTIF formula – that is, count the number of cells in column C that =1. You should see that 10 out of the 28 customers bought the special meal, or $p=0.3571$. Actually, note that since column C contains only 0's and 1's, we could have used =SUMIF and even just =SUM for column C and obtained the answer of 10. Furthermore, if we'd used =AVERAGE for column C we would have obtained the value of $p=0.3571$ directly – try it out!

To obtain a 95% confidence interval estimate for π , we use exactly the same process that we used to obtain a confidence interval for the mean amount spent, this time using the values in column C for the *Input Y Range*, and a column of 1's (create this in column D, and call it "Proportion") for the *Input X Range*, as follows:

And the output:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.597614					
R Square	0.357143					
Adjusted R Square	0.320106					
Standard Error	0.48795					
Observations	28					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	3.571429	3.571429	15	0.000651	
Residual	27	6.428571	0.238095			
Total	28	10				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
Proportion	0.357143	0.092214	3.872983	0.000619	0.167936	0.54635

Again, the "Lower 95%" and "Upper 95%" columns give us our 95% confidence interval for π : the two values tell us that we can be 95% sure that the true average proportion of customers who opted for the special meal is between 0.1679 and 0.5464.

N.B. A sample size of 28 is fairly small, so the shortcut approximation using the value from the standard normal distribution (1.96) to calculate the confidence interval is not particularly accurate in this case:

$$\text{Lower: } p - 1.96 \sqrt{\frac{p(1-p)}{n}} = 0.3571 - 1.96 \sqrt{\frac{0.3571(1-0.3571)}{28}} = 0.1797$$

$$\text{Upper: } p + 1.96 \sqrt{\frac{p(1-p)}{n}} = 0.3571 + 1.96 \sqrt{\frac{0.3571(1-0.3571)}{28}} = 0.5346$$

3. Confidence Interval for the Correlation

Now let's consider the correlation between amount spent by a customer and the age of the customer: *Sheet 3* provides this data.

First, calculate the correlation between amount spent and age of the customer using Excel's =CORREL function. You should obtain a value of 0.1551 – a weak positive correlation, meaning among customers of this fast food chain, older customers tend to spend more.

Before we can use the *Regression* Data Analysis tool, we need to scale the data by their respective standard deviations. First, calculate the sample standard deviation of the data in columns B and C in cells D2:E3 as shown below:

	A	B	C	D	E	F
1	Customer ID	Amount Spent \$	Customer Age	Correlation:	0.155064	
2	1	6.25	18	Std Dev Amount Spent:	0.869917	
3	2	7.8	26	Std Dev Age:	=STDEV.S(C2:C29)	
4	3	5.5	17			
5	4	6.75	24			
6	5	8.25	22			

Next create 2 new columns, "Amount Spent/Std. Dev. Amount Spent" and "Age/Std. Dev. Age", in columns F and G:

POISSON.... : X ✓ fx =C2/\$E\$3							
	A	B	C	D	E	F	G
1	Customer ID	Amount Spent \$	Customer Age	Correlation:	0.155064	Amount/Std Dev	Age/Std Dev
2	1	6.25	18	Std Dev Amount Spent:	0.869917	7.185	=C2/\$E\$3
3	2	7.8	26	Std Dev Age:	12.66682	8.966	2.053
4	3	5.5	17			6.322	1.342
5	4	6.75	24			7.759	1.895
6	5	8.25	22			9.484	1.737

Now use column F as the *Input Y Range* and column G as the *Input X Range* in the dialog box for the *Regression* Data Analysis tool. Be sure to un-tick the *Constant is Zero* checkbox:

Regression ? X

Input

Input Y Range: ↑

Input X Range: ↑

☒ Labels ☐ Constant is Zero

☒ Confidence Level: %

Output options

☒ Output Range: ↑

☐ New Worksheet Ply:

☐ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK Cancel Help

You should obtain the following output:

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0.155064					
R Square	0.024045					
Adjusted R Square	-0.01349					
Standard Error	1.006723					
Observations	28					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	0.649211	0.649211	0.640568	0.430758	
Residual	26	26.35079	1.013492			
Total	27	27				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	7.556043	0.471125	16.03829	5.33E-15	6.587631	8.524455
Age/Std Dev	0.155064	0.193744	0.800355	0.430758	-0.24318	0.553311

Here we see the confidence interval for correlation is (-0.2432, 0.5533) – very wide, and indicates that the true correlation between customer age and amount spent could be a positive, negative or potentially zero value. That is, we can't be particularly confident in there even being a correlation between age and amount spent by customers! This is an example of why a point estimate from a sample (e.g. the correlation coefficient of 0.1551) needs to be assessed alongside measures that take into account the degree of uncertainty inherent in the sample (e.g. the confidence interval).