

MODULE 5 LIVE LECTURE

INFERENTIAL STATISTICS AND ANALYTICS

INFERENCEAL STATISTICS AND ANALYTICS–MODULE 5 LIVE LECTURE

- ▶ Module 5 Linear Regression
- ▶ Module 5 homework
- ▶ Quiz 5
- ▶ Course project phase 5
- ▶ Module 5 Live classroom Grading
- ▶ Summary

INFERENCEAL STATISTICS AND ANALYTICS–MODULE 5 LIVE LECTURE

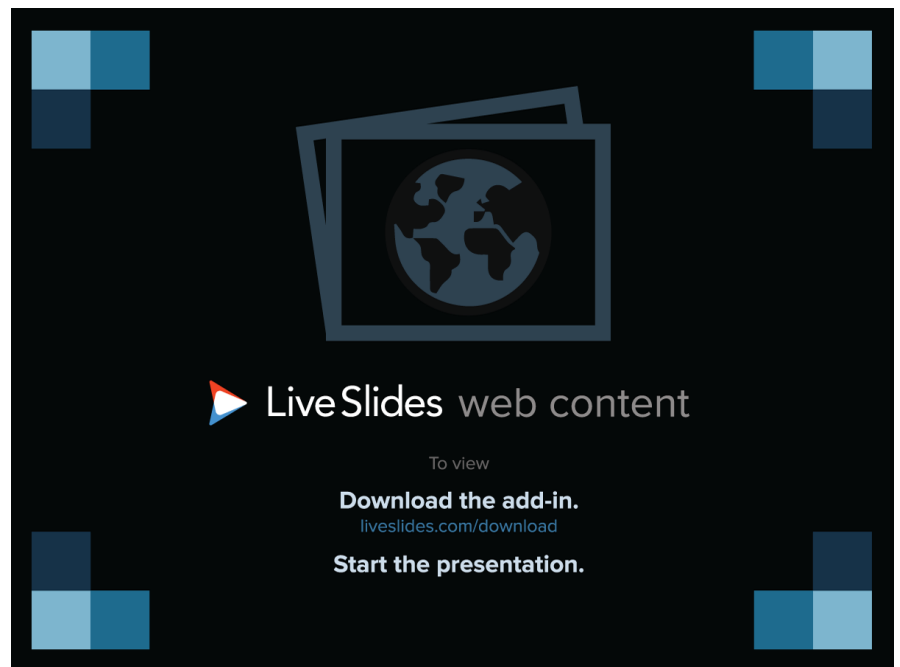
- ▶ ***Module 5 Linear Regression***

- ▶ Module 5 homework
- ▶ Quiz 5
- ▶ Course project phase 5
- ▶ Module 5 Live classroom Grading
- ▶ Summary

MODULE 5 LINEAR REGRESSION

Linear regression is a statistical method that allows us to summarize and study relationships between two variables.

- ▶ The linear correlation coefficient, r , can take a **range of values from -1 to +1**.
- ▶ A value of 0 indicates that there is no association between the two variables.
- ▶ An r of -1 indicates a perfect negative linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables.



<https://www.youtube.com/watch?v=KsVBBJRb9TE>

INFERENCEAL STATISTICS AND ANALYTICS–MODULE 5 LIVE LECTURE

- ▶ ***Module 5 Linear Regression***
- ▶ ***Module 5 homework***
- ▶ Quiz 5
- ▶ Course project phase 5
- ▶ Module 5 Live classroom Grading
- ▶ Summary

MODULE 5 HOMEWORK

1. Using the paired height/pulse data for males, we get the regression equation: where x represents height (cm) and the pulse rate is in beats per minute.

$$\hat{y} = 73.9 + 0.023x$$

\hat{y} represents the predicted pulse rate.

Predictor variable

The predictor variable represents the height. It explains variations in the response variable; in an experimental study, it is manipulated by the researcher.

Response Variable

The response variable represents the pulse rate. The response variable is also known as the dependent or outcome variable, its value is predicted or its variation is explained by the predictor variable; in an experimental study, this is the outcome that is measured following manipulation of the explanatory variable

MODULE 5 HOMEWORK

2. The data below represents the weight and price for 10 randomly selected pieces of gold.

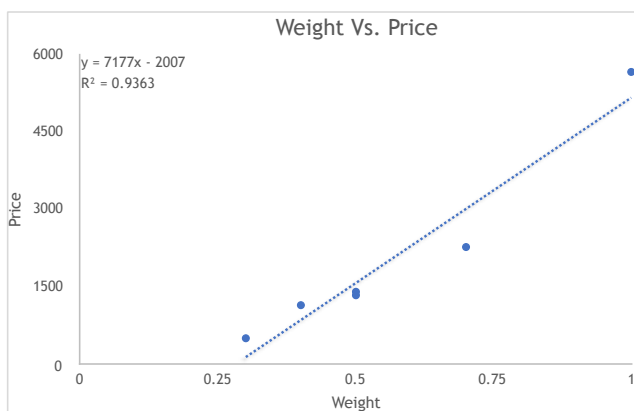
Weight (Ounces)	Price (Dollars)
0.3	510
0.4	1151
0.5	1343
0.5	1410
1.0	5669
0.7	2277

There appears to be a positive correlation between weight and price of gold.

$$n=6$$

$$r^2 = 0.9363$$

$$y=7177x-2007$$



$$r^2 = 0.9363 \rightarrow r=0.968$$

$$y=7177x-2007$$

eg. $x=1.5$ ounces

$$y=7177x1.5-2007$$

MODULE 5 HOMEWORK

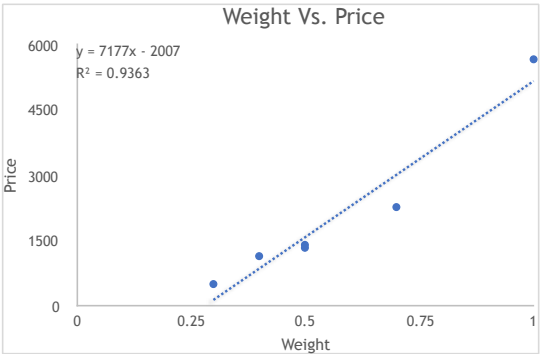
2.The data below represents the weight and price for 10 randomly selected pieces of gold.

Table of Critical Values: Pearson Correlation

Degrees of Freedom = $N-2=6-2=4$

df	0.1	0.05	0.01
1	0.988	0.997	0.999
2	0.900	0.950	0.990
3	0.805	0.878	0.959
4	0.729	0.811	0.917

...



Compare $r = 0.968$ to the critical values of the correlation coefficient

For $n = 6$ and $\alpha = .05$, the critical values are $r = \pm 0.811$

Since $|0.968| > 0.811$, we conclude that there is sufficient evidence to support the claim

MODULE 5 HOMEWORK

3. You are given a correlation coefficient of $r = 0.933$ where x = weight of males and y = the waist size of males.

The coefficient of determination $r^2 = 0.933 \times 0.933$

The coefficient of determination r^2 is the amount of the variation in y that is explained by the regression line. It is the ratio of the explained variation to the total variation.

$1 - r^2$ is the amount of the unexplained variation. It is the ratio of the unexplained variation to the total variation. The variation in waist size is explained by other factors.

MODULE 5 HOMEWORK

4. Let the predictor variable x represent the heights (cm) of females and let the response variable y represent the weights (kg) of females. A sample of 40 heights and weights result in a standard error = 17.5436.

$$s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}$$

The value of 17.5436 is the standard error of the estimate which is the measure of the differences in the observed weights and the predicted weights from the regression equation. It is a measure of the sample points about the regression line.

MODULE 5 HOMEWORK

5. An investigator analyzed the leading digits of the amounts from 200 checks issued by three suspect companies. The frequencies were found to be 68, 40, 18, 19, 8, 20, 6, 9, 12 and those digits correspond to the leading digits of 1, 2, 3, 4, 5, 6, 7, 8, and 9, respectively. If the observed frequencies are substantially different from the frequencies expected with Benford's law, the check amounts appear to be the result of fraud. Use a 0.05 significance level to test for goodness-of-fit with Benford's law.

d	O	E	O-E	(O-E)^2	(O-E)^2/E
1	68	200x30.1%=60.2	68-60.2=7.8	7.8x7.8	a1=7.8x7.8/60.2
2	40	200x17.6%=35.2	40-35.2=4.8	4.8x4.8	a2=0.655
3	18	200x12.5%=25	18-25=-7	7x7	a3=1.96
4	19	200x9.7%=19.4	19-19.4=-0.4	0.4x0.4	a4=0.008
5	8	200x7.9%=15.8	8-15.8=-7.8	7.8x7.8	a5=3.851
6	20	200x6.7%=13.4	20-13.4=6.6	6.6x6.6	a6=3.251
7	6	200x5.8%=11.6	6-11.6=-5.6	5.6x5.6	a7=2.703
8	9	200x5.1%=10.2	9-10.2=-1.2	1.2x1.2	a8=0.141
9	12	200x4.6%=9.2	12-9.2=2.8	2.8x2.8	a9=0.852
total	200	200			sum =14.432

a. Calculate the χ^2 test statistic

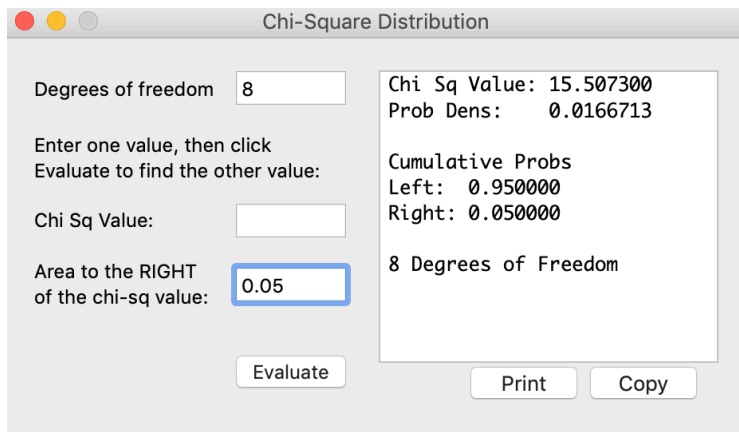
$$\chi^2 = \sum \frac{(E - o)^2}{E} = a1 + a2 + a3 + a4 + a5 + a6 + a7 + a8 + a9 = 14.432$$

MODULE 5 HOMEWORK

b. Calculate the χ^2 critical value.

Use a 0.05 significance level to test for goodness-of-fit with Benford's law.

$$Df = 9 - 1 = 8$$



A screenshot of a web-based calculator titled "Chi-Square Distribution". The interface is divided into two main sections. The left section contains input fields: "Degrees of freedom" with the value "8", "Chi Sq Value:" with an empty field, and "Area to the RIGHT of the chi-sq value:" with the value "0.05". Below these is an "Evaluate" button. The right section displays the results: "Chi Sq Value: 15.507300", "Prob Dens: 0.0166713", "Cumulative Probs" with "Left: 0.950000" and "Right: 0.050000", and "8 Degrees of Freedom". At the bottom right of the results area are "Print" and "Copy" buttons.

Input	Output
Degrees of freedom: 8	Chi Sq Value: 15.507300
Area to the RIGHT of the chi-sq value: 0.05	Prob Dens: 0.0166713
	Cumulative Probs
	Left: 0.950000
	Right: 0.050000
	8 Degrees of Freedom

MODULE 5 HOMEWORK

- Calculate the χ^2 test statistic.
- Calculate the χ^2 critical value.

Goodness-of-Fit: Unequal E

Significance:

Enter Expected Frequencies

☐ As Proportions

☒ As Counts

Observed Column:

Expected Column:

Evaluate Plot

Clear Copy Paste

Row	1	2
Num Categories: 1	68	60.2
Degrees of freedom: 2	40	35.2
3	18	25
Test Statistic: 4	19	19.4
Critical χ^2 : 5	8	15.8
P-Value: 0.006	20	13.4
7	6	11.6
8	9	10.2
9	12	9.2
10		
11		
12		
13		

Print Copy

INFERENCEAL STATISTICS AND ANALYTICS–MODULE 5 LIVE LECTURE

- ▶ ***Module 5 Linear Regression***
- ▶ ***Module 5 homework***
- ▶ ***Quiz 5***
- ▶ Course project phase 5
- ▶ Module 5 Live classroom Grading
- ▶ Summary

QUIZ 5

1). Given the linear correlation coefficient r and the sample size n , determine the critical values of r and use your finding to state whether or not the given r represents a significant linear correlation. Use a significance level of 0.05. $r = 0.543$, $n = 25$ $df = n - 2 = 25 - 2 = 23$ 0.05 significance level

Table of Critical Values: Pearson Correlation

Degrees of Freedom = $N - 2 = 25 - 2 = 23$

df	0.1	0.05	0.01
1	0.988	0.997	0.999
2	0.900	0.950	0.990
...			
22	0.344	0.404	0.515
23	0.337	0.396	0.505
24	0.330	0.388	0.496
25	0.323	0.381	0.487

$r = 0.543$

Since $|r| >$ critical value 0.396,

we can reject the null hypothesis and conclude that there is sufficient evidence to support the claim of a linear correlation. That is significant linear correlation

QUIZ 5

2). Find the coefficient of determination, given that the value of the linear correlation coefficient, r , is 0.756.

The coefficient of determination = $r^2 = 0.756 \times 0.756$

QUIZ 5

3). Ten pairs of data yield $r = 0.003$ and the regression equation $y = 2 + 3x$. Also, the mean $y = 5.0$. What is the best predicted value of y for $x = 2$?

Table of Critical Values: Pearson Correlation

Degrees of Freedom = $N-2=10-2=8$

df	0.1	0.05	0.01
1	0.988	0.997	0.999
2	0.900	0.950	0.990
...			
7	0.584	0.666	0.798
8	0.549	0.632	0.765
9	0.521	0.602	0.735
10	0.497	0.576	0.708
11	0.476	0.553	0.684

Compare $r = 0.003$ to the critical values of the correlation coefficient for $\alpha = .05$ and $\alpha = .01$ $df=10-2=8$

For $n = 10$ and $\alpha = .05$, the critical values are $r = \pm 0.632$

Since $|r| = 0.003$ is less than the critical values 0.632 , there is no linear correlation.

Therefore, the regression equation is NOT a good model.

Since the regression equation is NOT a good model, use sample mean to find the predicted value of y .

The predicted value of y for $x=2$ is the mean $y=5.0$

QUIZ 5

4) The linear correlation coefficient, r , between systolic and diastolic blood pressure readings is 0.585 for a sample of 5 patients. Using a significance level of 0.05, what should you conclude

Table of Critical Values: Pearson Correlation

Degrees of Freedom = $N-2=5-2=3$

df	0.1	0.05	0.01
1	0.988	0.997	0.999
2	0.900	0.950	0.990
3	0.805	0.878	0.959
4	0.729	0.811	0.917

...

Compare $r = 0.585$ to the critical values of the correlation coefficient

For $n = 5$ and $\alpha = .05$, the critical values are $r = \pm 0.878$

Since $|r| = 0.585$ is less than the critical values 0.878, there is insufficient evidence to support the claim that there is a linear correlation

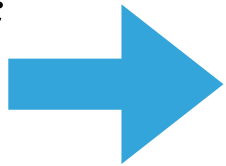
QUIZ 5

5). Use the given data to find the equation of the regression line. Round the final answer to three significant digits, if necessary.

x 6 8 20 28 36

y 2 4 13 20 30

$$\hat{y} = b_0 + b_1x$$



Correlation and Regression

Significance:

Select the columns to be used for the x and y variables.

x variable column y variable column

Sample size, n: 5
Degrees of freedom: 3

Correlation Results:
Correlation coeff, r: 0.907
Critical r: ±0.818
P-value (two-tailed): 0.000

Regression Results:
Y= b0 + b1x:
Y Intercept, b0: 0.571
Slope, b1: 0.857

Total Variation: 140
Explained Variation: 102.86
Unexplained Variation: 37.14
Standard Error: 6.09
Coeff of Det, R^2: 0.818

Row	1	2
1	6	2
2	8	4
3	20	13
4	28	20
5	36	30

QUIZ 5

6) A regression equation is obtained for a collection of paired data. It is found that the total variation is 20.711, the explained variation is 18.592, and the unexplained variation is 2.119. Find the coefficient of determination.

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

Total variation = Explained variation + Unexplained variation

$$r^2 = \frac{18.592}{18.592 + 2.119}$$

QUIZ 5

The paired data below consist of the temperatures on randomly chosen days and the amount a certain kind of plant grew (in millimeters):

Temp 62 76 50 51 71 46 51 44 79

Growth 36 39 50 13 33 33 17 6 16

7) Find the value of the linear correlation coefficient r .

8) Find the value of the coefficient of determination r^2

Correlation and Regression

Significance:

Select the columns to be used for the x and y variables.

x variable column y variable column

Sample size, n: 9
Degrees of freedom: 7

Correlation Results:
Correlation coeff, r:
Critical r:
P-value (two-tailed):

Regression Results:
Y= $b_0 + b_1x$:
Y Intercept, b_0 :
Slope, b_1 :

Total Variation:
Explained Variation:
Unexplained Variation:
Standard Error:
Coeff of Det, R^2 :

	1	2
ow	62	36
	76	39
	50	50
	51	13
	71	33
	46	33
	51	17
	44	6
	79	16
0		
1		

QUIZ 5

9) A survey was administered to 1005 students who were asked which day of the week was best for working on homework assignments. The results are as follows:

Sun	Mon	Tue	Wed	Thu	Fri	Sat
382	20	9	19	41	11	523

A **goodness-of-fit test** is used to test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution.

The number of degrees of freedom for a goodness of fit test is simply one less than the number of levels of our variable. Since $n=7$, we have $n - 1$ degrees of freedom.

QUIZ 5

10)A survey was administered to 1005 students who were asked which day of the week was best for working on homework assignments. The results are as follows:

If we wanted to test the claim using a goodness-of-fit test, what is the critical χ^2 value assuming a 0.05 significance level?

Sun	Mon	Tue	Wed	Thu	Fri	Sat
382	20	9	19	41	11	523

Goodness-of-Fit: Equal Exp.

Significance: 0.05

Select a column to be the Observed Frequencies:

1

Evaluate

Plot

Num Categories: 7

Degrees of freedom: 6

Expected Freq: 143.57

Test Statistic, χ^2 : 143.57

Critical χ^2 : 12.59

P-Value: 0.0000

Clear

Copy

Row	1	
1	382	
2	20	
3	9	
4	19	
5	41	
6	11	
7	523	
8		
9		
10		
11		

INFERENCEAL STATISTICS AND ANALYTICS–MODULE 5 LIVE LECTURE

- ▶ ***Module 5 Linear Regression***
- ▶ ***Module 5 homework***
- ▶ ***Quiz 5***
- ▶ ***Course project phase 5***
- ▶ Module 5 Live classroom Grading
- ▶ Summary

COURSE PROJECT PHASE 5

- ▶ This week you will submit Phase 5, the final phase, of your course project. For Phase 5 of your course project, you will want to review grade feedback from your Phase 4 submission and make any necessary corrections.
- ▶ Once you have made your corrections, you will make your final submission for the course project.

INFERENCEAL STATISTICS AND ANALYTICS–MODULE 5 LIVE LECTURE

- ▶ ***Module 5 Linear Regression***
- ▶ ***Module 5 homework***
- ▶ ***Quiz 5***
- ▶ ***Course project phase 5***
- ▶ ***Module 5 Live classroom Grading***
- ▶ Summary

MODULE 5 LIVE CLASSROOM GRADING

The live classroom session archive as a URL in the course will be added after the session has ended. The following is how you will receive your points for the module 5 live classroom session:

Question:

What is the range of values for the linear correlation coefficient r ?

Please go to [module 5 live classroom](#), enter your response to receive your points.

You have until midnight CST on Sunday to confirm that you have viewed the live classroom session archive.

INFERENCEAL STATISTICS AND ANALYTICS–MODULE 5 LIVE LECTURE

- ▶ ***Module 5 Linear Regression***
- ▶ ***Module 5 homework***
- ▶ ***Quiz 5***
- ▶ ***Course project phase 5***
- ▶ ***Module 5 Live classroom Grading***
- ▶ ***Summary***

SUMMARY

- ▶ The linear correlation coefficient, r , can take a **range of values from -1 to +1**.
- ▶ A value of 0 indicates that there is no association between the two variables.
- ▶ An r of -1 indicates a perfect negative linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables.
- ▶ Module 5 Question: What is the range of values for the linear correlation coefficient r ?