



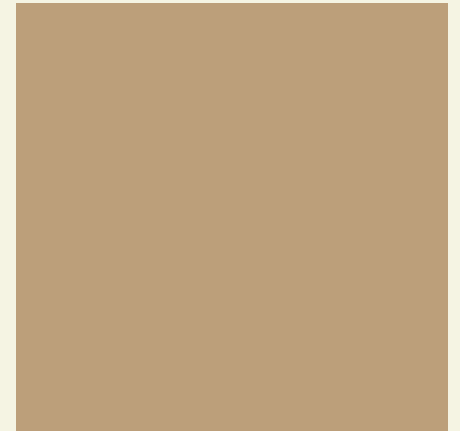
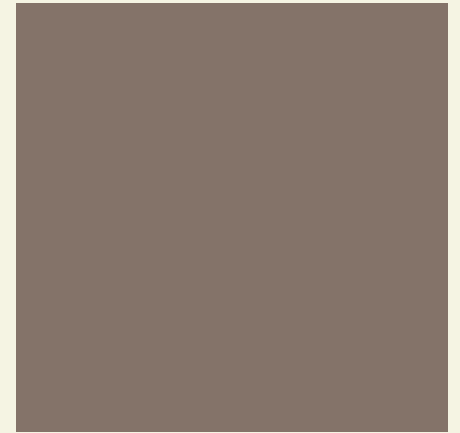
ICCPP-STATISTICS

- Testing the Significance of Correlation Coefficient

Vishal Lohchab

*Scientific Assistant of
Prof. Dr. Hans-Werner Gessmann
Director ICCPP International*



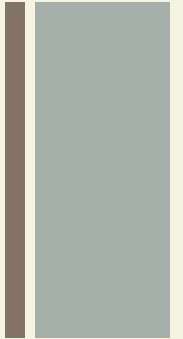


Francis Galton (1822-1911)

Testing the Significance of
Correlation Coefficient



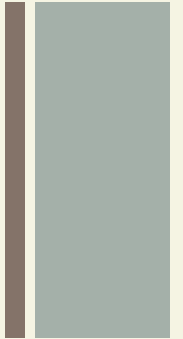
Testing the Significance of Correlation Coefficient



- The correlation coefficient r tells us about the strength and direction of the linear relationship between x and y .
- However, the reliability of the linear model also depends on how many observed data points are in the sample.



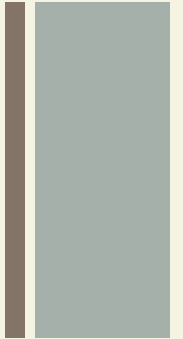
Testing the Significance of Correlation Coefficient



- We need to look at both the value of the correlation coefficient r and the sample size n together.
- We perform a hypothesis test of the "significance of the correlation coefficient" to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.



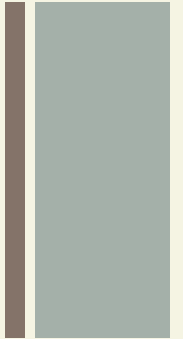
Testing the Significance of Correlation Coefficient



- The sample data is used to compute r , the correlation coefficient for the sample.
- If we had data for the entire population, we could find the population correlation coefficient.



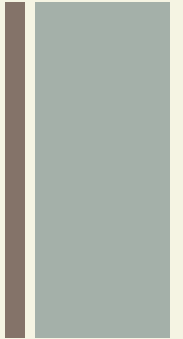
Testing the Significance of Correlation Coefficient



- But because we only have sample data, we can not calculate the population correlation coefficient.
- The sample correlation coefficient r is our estimate of the unknown population correlation coefficient.



Testing the Significance of Correlation Coefficient

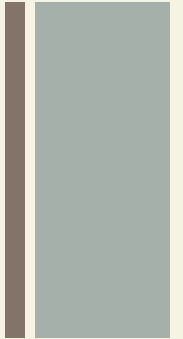


- The symbol for the population correlation coefficient is ρ , the Greek letter "rho".

ρ = population correlation coefficient (unknown)

r = sample correlation coefficient (known; calculated from sample data)

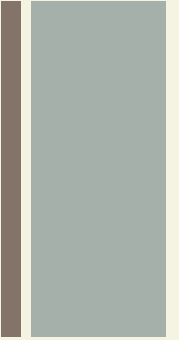
+ Use



- The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to 0" or "significantly different from 0".
- We decide this based on the sample correlation coefficient r and the sample size n .

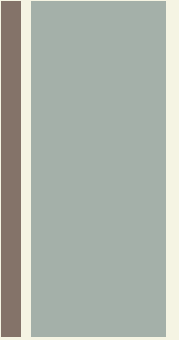
+ Significant

- If the test concludes that the correlation coefficient is significantly different from 0, we say that the correlation coefficient is “significant”.



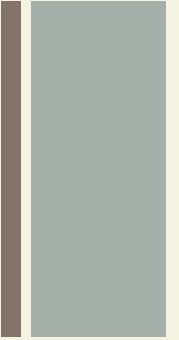
+ Significant

- Conclusion: “There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from 0.”



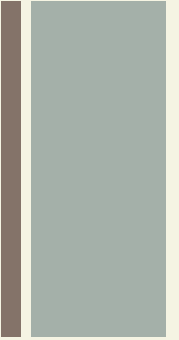
+ Significant

- What the conclusion means: There is a significant linear relationship between x and y . We can use the regression line to model the linear relationship between x and y in the population.



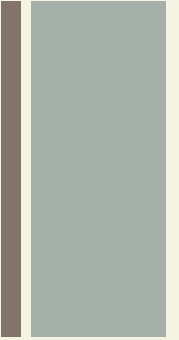
+ Not significant

- If the test concludes that the correlation coefficient is not significantly different from 0 (it is close to 0), we say that correlation coefficient is "not significant".



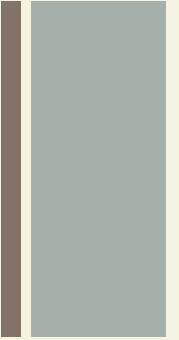
+ Not significant

- Conclusion: “There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from 0.”

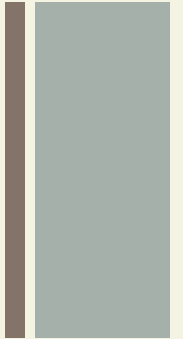


+ Not significant

- What the conclusion means: There is not a significant linear relationship between x and y . Therefore we can not use the regression line to model a linear relationship between x and y in the population.



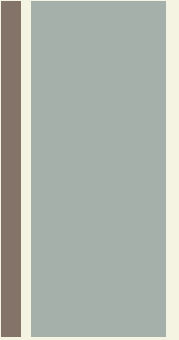
+ Note



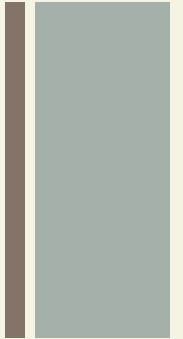
- If r is significant and the scatter plot shows a linear trend, the line can be used to predict the value of y for values of x that are within the domain of observed x values.
- If r is not significant or if the scatter plot does not show a linear trend, the line should not be used for prediction.

+ Note

- If r is significant and if the scatter plot shows a linear trend, the line may not be appropriate or reliable for prediction outside the domain of observed x values in the data.



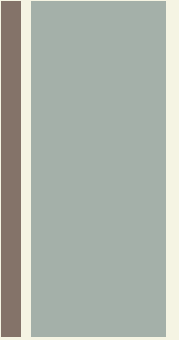
+ Hypothesis Test



- Null Hypothesis $H_0 : \rho = 0$
- Alternate Hypothesis $H_a : \rho \neq 0$

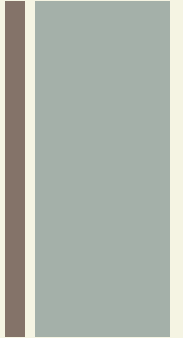
+ Null Hypothesis Means

- Null Hypothesis H_0 : The population correlation coefficient is not significantly different from zero.
- There is not a significant linear relationship (correlation) between x and y in the population.





Alternate Hypothesis Means



- Alternate Hypothesis H_a : The population correlation coefficient is significantly different from zero.
- There is a significant linear relationship (correlation) between x and y in the population.

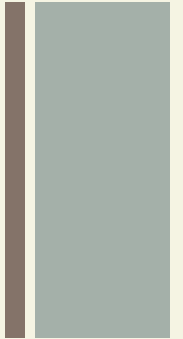
+ Drawing a Conclusion

- There are two methods to make the decision. Both methods are equivalent and give the same result.

METHOD 1: Using a p -value to make a decision.

METHOD 2: Using a table of Critical Values to make a decision.

+ Note



- Using the p-value method, you could choose any appropriate significance level you want; you are not limited to using $\alpha = 0.05$.
- If you wanted to use a different significance level than 5% with the critical value method, you would need different tables of critical values.

+

Method 1

Using a p-value to make a decision

- To calculate the p-value using Lin Reg T TEST:

On the Lin Reg T TEST input screen, on the line prompt for β or ρ , highlight “ $\neq 0$ ”.

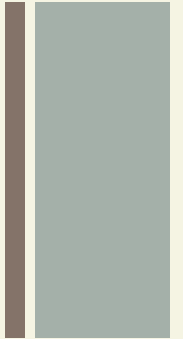
The output screen shows the p-value on the line that reads “p=”.

- (Most computer statistical software can calculate the p-value.)

+

Method 1

Using a p-value to make a decision



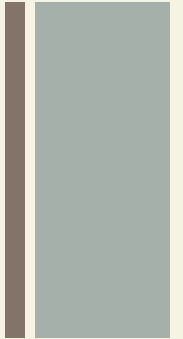
- If the p-value is less than the significance level
($\alpha = 0.05$)

Decision: Reject the null hypothesis.

- If the p-value is not less than the significance level
($\alpha = 0.05$)

Decision: do not reject the null hypothesis.

+ Calculation Notes



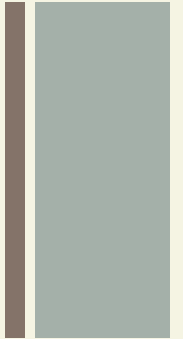
- You will use technology to calculate the p-value.
- Way to calculate the p-value (p) given by Lin Reg T Test is the command $2 * tcdf(\text{abs}(t), 10^{99}, n-2)$ in 2nd DISTR.



Method 2

Using a table of Critical Values to make a decision

- The 95% Critical Values of the Sample Correlation Coefficient Table can be used to give you a good idea of whether the computed value of r is significant or not.
- Compare r to the appropriate critical value in the table.

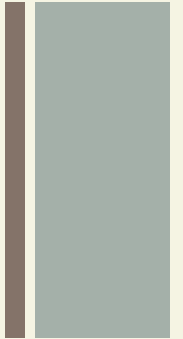




Method 2

Using a table of Critical Values to make a decision

- If r is not between the positive and negative critical values, then the correlation coefficient is significant.
- If r is significant, then you may want to use the line for prediction.



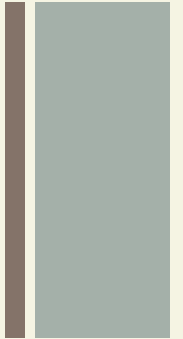
+ Assumptions

- There is a linear relationship in the population that models the average value of y for varying values of x .

In other words, the expected value of y for each particular value lies on a straight line in the population.



Assumptions



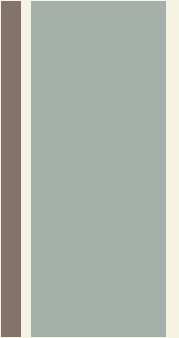
- The y values for any particular x value are normally distributed about the line. This implies that there are more y values scattered closer to the line than are scattered farther away.

Assumption (1) implies that these normal distributions are centered on the line: the means of these normal distributions of y values lie on the line.

+ Assumptions

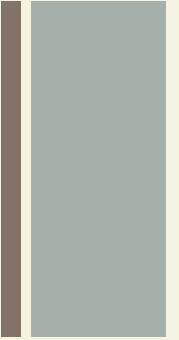
- The standard deviations of the population y values about the line are equal for each value of x .

In other words, each of these normal distributions of y values has the same shape and spread about the line.



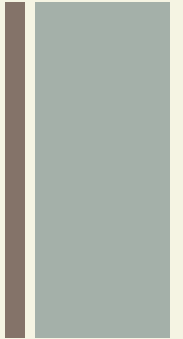
+ Assumptions

- The residual errors are mutually independent (no pattern).
- The data are produced from a well-designed, random sample or randomized experiment.





References



Illowsky, B., & Dean, S. (2021, November 6). Testing the Significance of the Correlation Coefficient. De Anza College. <https://stats.libretexts.org/@go/page/800>, date 15.11.21, 16:30 h MET

https://en.wikipedia.org/wiki/Francis_Galton, date 15.11.21, 16:00 h MET

<https://www.opentextbooks.org.hk/ditatopic/9498>, date 15.11.21, 17:00 h MET