Climate and Oceans Support
Program in the Pacific

ACCESS-S Workshop

MODULE: Statistics Fundamentals

Climate and Oceans Support Program in the Pacific

Topics in this module

- Mean / average
- Anomalies
- Median
- Quantiles and distributions
- Correlation
- Statistically significant correlation
- Root Mean Square Error

Expected learning outcomes

- Understanding the mean, median and anomalies and how to calculate them
- Understanding distributions and their relationship to probability
- Understanding how to calculate statistical significance for trend lines and correlations

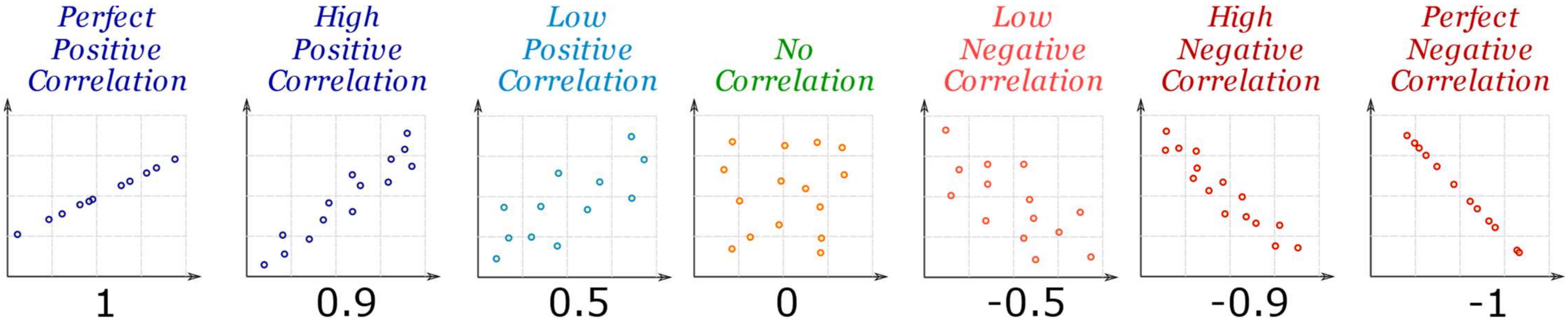
Climate and Oceans Support
Program in the Pacific

Correlation – how related are two variables?

- We test a **dependent** variable's relationship with an **independent** variable
- Example: Rainfall (dependent); ENSO (independent)
- The **Correlation Coefficient (r)** measures the strength of the relationship – it can vary between –1 and +1
- Values of –1 and +1 are **perfect** in which all the observations lie on a straight line
- **Positive** correlation: dependent variable increases as the independent variable increases
- **Negative** correlation: dependent variable decreases as the independent variable increases
- **r** relates to the **scatter** of observations about the regression line of best fit
- Correlation does **not** imply Causation
- Correlation can be used to calculate model skill

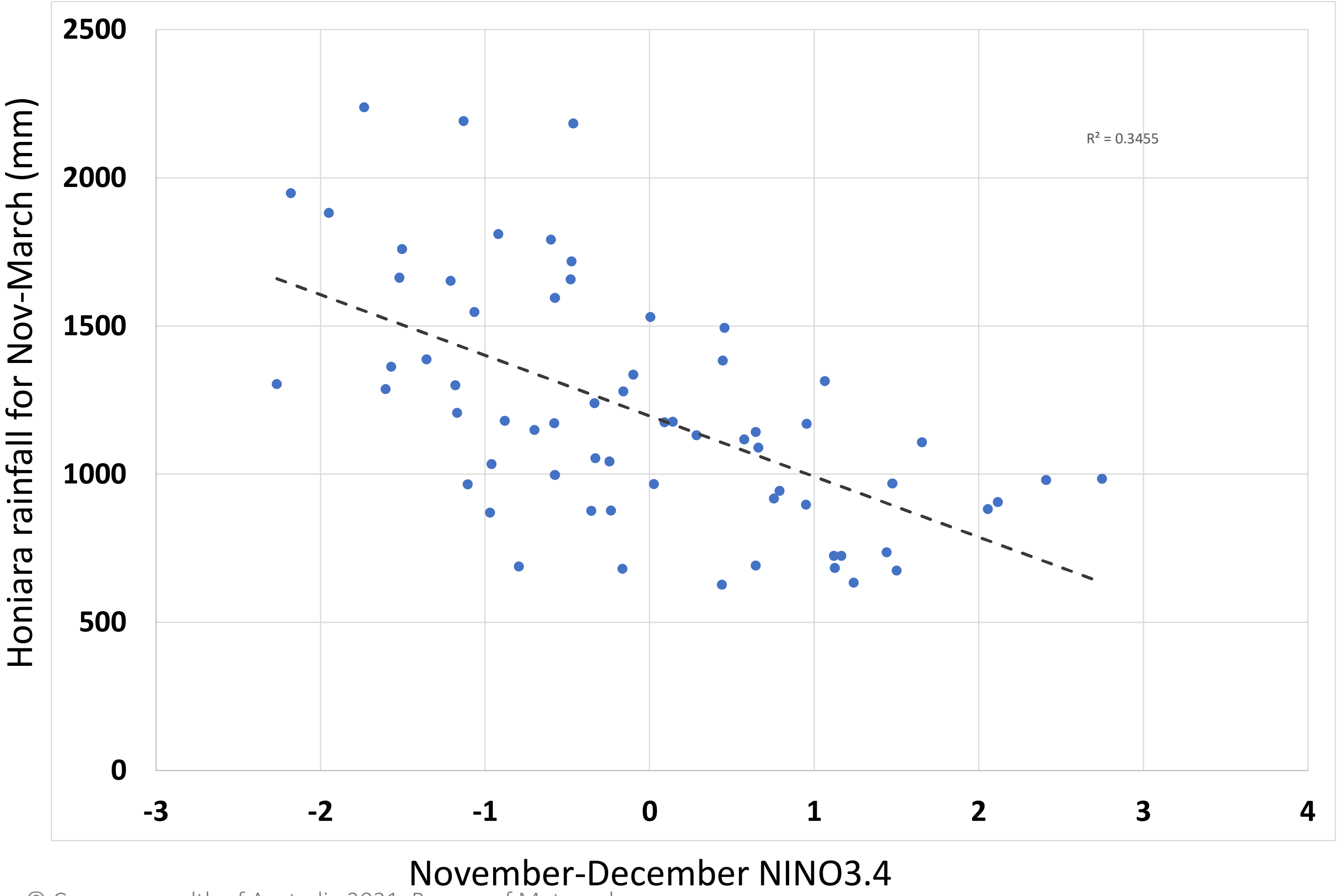
Climate and Oceans Support Program in the Pacific

Correlation coefficient examples



The number below each graph is the value of **r**, the correlation coefficient

Using excel to calculate the correlation



What is the correlation between ENSO and rainfall at Honiara?

Use Excel to test the correlation of November to March Honiara rainfall compared to the November to December NINO3.4.

To calculate the correlation coefficient between two arrays of numbers: the formula is **=Correl(array1:array2)**

There is a negative correlation between the Nov-Dec value of NINO3.4 and the total Nov-March rainfall at Honiara.

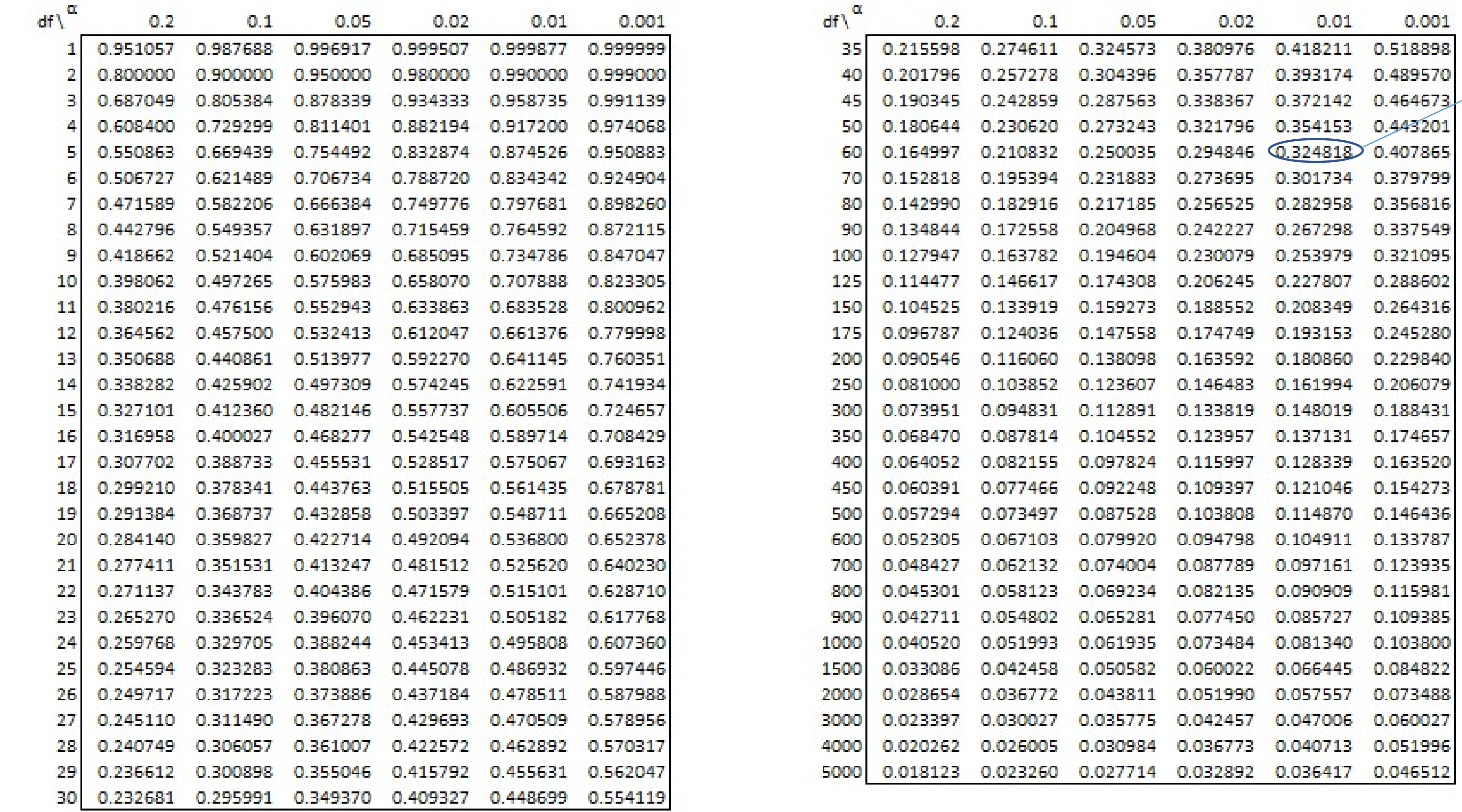
Excel calculates –0.58782 for **r**. If we square this, we get the **R^2** value shown on the graph.

But is it **statistically significant**?

Using significance tables for correlation significance

The correlation coefficient (**r**) can be checked against a table of **critical r** values for **different levels of significance** [e.g. 0.05 (5%) or 0.01 (1%)] and **degrees of freedom (df)**

degrees of freedom (df) = n – 1, where n is the number of points on the graph, i.e. the sample size



df\ α	0.2	0.1	0.05	0.02	0.01	0.001
1	0.951057	0.987688	0.996917	0.999507	0.999877	0.999999
2	0.800000	0.900000	0.950000	0.980000	0.990000	0.999000
3	0.687049	0.805384	0.878339	0.934333	0.958735	0.991139
4	0.608400	0.729299	0.811401	0.882194	0.917200	0.974068
5	0.550863	0.669439	0.754492	0.832874	0.874526	0.950883
6	0.506727	0.621489	0.706734	0.788720	0.834342	0.924904
7	0.471589	0.582206	0.666384	0.749776	0.797681	0.898260
8	0.442796	0.549357	0.631897	0.715459	0.764592	0.872115
9	0.418662	0.521404	0.602069	0.685095	0.734786	0.847047
10	0.398062	0.497265	0.575983	0.658070	0.707888	0.823305
11	0.380216	0.476156	0.552943	0.633863	0.683528	0.800962
12	0.364562	0.457500	0.532413	0.612047	0.661376	0.779998
13	0.350688	0.440861	0.513977	0.592270	0.641145	0.760351
14	0.338282	0.425902	0.497309	0.574245	0.622591	0.741934
15	0.327101	0.412360	0.482146	0.557737	0.605506	0.724657
16	0.316958	0.400027	0.468277	0.542548	0.589714	0.708429
17	0.307702	0.388733	0.455531	0.528517	0.575067	0.693163
18	0.299210	0.378341	0.443763	0.515505	0.561435	0.678781
19	0.291384	0.368737	0.432858	0.503397	0.548711	0.665208
20	0.284140	0.359827	0.422714	0.492094	0.536800	0.652378
21	0.277411	0.351531	0.413247	0.481512	0.525620	0.640230
22	0.271137	0.343783	0.404386	0.471579	0.515101	0.628710
23	0.265270	0.336524	0.396070	0.462231	0.505182	0.617768
24	0.259768	0.329705	0.388244	0.453413	0.495808	0.607360
25	0.254594	0.323283	0.380863	0.445078	0.486932	0.597446
26	0.249717	0.317223	0.373886	0.437184	0.478511	0.587988
27	0.245110	0.311490	0.367278	0.429693	0.470509	0.578956
28	0.240749	0.306057	0.361007	0.422572	0.462892	0.570317
29	0.236612	0.300898	0.355046	0.415792	0.455631	0.562047
30	0.232681	0.295991	0.349370	0.409327	0.448699	0.554119

df\ α	0.2	0.1	0.05	0.02	0.01	0.001
35	0.215598	0.274611	0.324573	0.380976	0.418211	0.518898
40	0.201796	0.257278	0.304396	0.357787	0.393174	0.489570
45	0.190345	0.242859	0.287563	0.338367	0.372142	0.464673
50	0.180644	0.230620	0.273243	0.321796	0.354153	0.443201
60	0.164997	0.210832	0.250035	0.294846	0.324818	0.407865
70	0.152818	0.195394	0.231883	0.273695	0.301734	0.379799
80	0.142990	0.182916	0.217185	0.256525	0.282958	0.356816
90	0.134844	0.172558	0.204968	0.242227	0.267298	0.337549
100	0.127947	0.163782	0.194604	0.230079	0.253979	0.321095
125	0.114477	0.146617	0.174308	0.206245	0.227807	0.288602
150	0.104525	0.133919	0.159273	0.188552	0.208349	0.264316
175	0.096787	0.124036	0.147558	0.174749	0.193153	0.245280
200	0.090546	0.116060	0.138098	0.163592	0.180860	0.229840
250	0.081000	0.103852	0.123607	0.146483	0.161994	0.206079
300	0.073951	0.094831	0.112891	0.133819	0.148019	0.188431
350	0.068470	0.087814	0.104552	0.123957	0.137131	0.174657
400	0.064052	0.082155	0.097824	0.115997	0.128339	0.163520
450	0.060391	0.077466	0.092248	0.109397	0.121046	0.154273
500	0.057294	0.073497	0.087528	0.103808	0.114870	0.146436
600	0.052305	0.067103	0.079920	0.094798	0.104911	0.133787
700	0.048427	0.062132	0.074004	0.087789	0.097161	0.123935
800	0.045301	0.058123	0.069234	0.082135	0.090909	0.115981
900	0.042711	0.054802	0.065281	0.077450	0.085727	0.109385
1000	0.040520	0.051993	0.061935	0.073484	0.081340	0.103800
1500	0.033086	0.042458	0.050582	0.060022	0.066445	0.084822
2000	0.028654	0.036772	0.043811	0.051990	0.057557	0.073488
3000	0.023397	0.030027	0.035775	0.042457	0.047006	0.060027
4000	0.020262	0.026005	0.030984	0.036773	0.040713	0.051996
5000	0.018123	0.023260	0.027714	0.032892	0.036417	0.046512

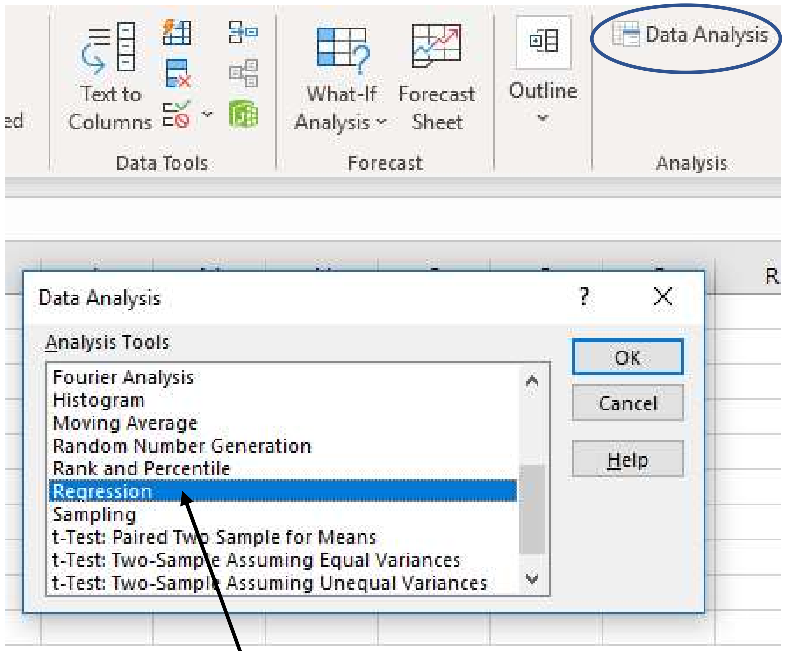
For example, a sample with 60 degrees of freedom needs a correlation of at least 0.3248 (positive or negative) to be significant at the 1% level. Significant at the 1% level is a high level of confidence.

Tables of critical r values can be found on the internet.

Climate and Oceans Support Program in the Pacific

How to calculate if the correlation is significant

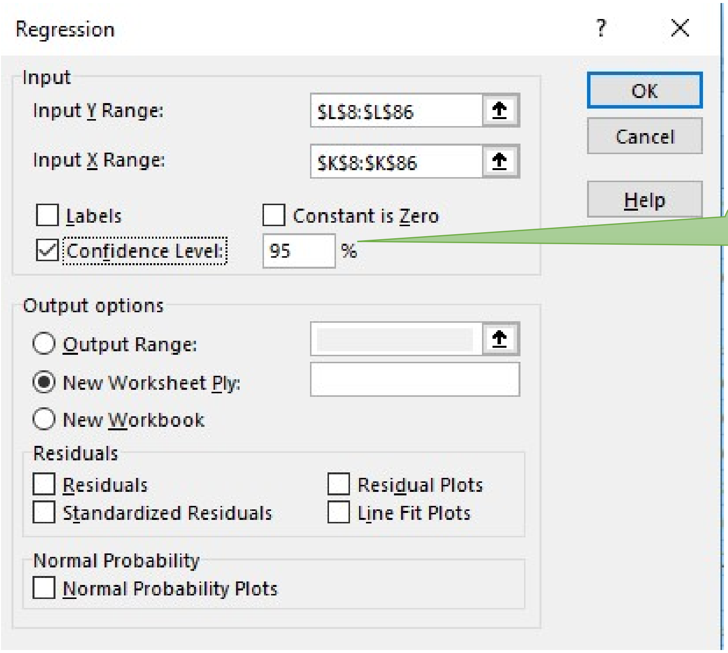
Use **Data Analysis in Excel**



Data Analysis ? ✕

Analysis Tools

- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test: Paired Two Sample for Means
- t-Test: Two-Sample Assuming Equal Variances
- t-Test: Two-Sample Assuming Unequal Variances



Regression ? ✕

Input

Input Y Range: $L$8:$L$86

Input X Range: $K$8:$K$86

□ Labels □ Constant is Zero

☑ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

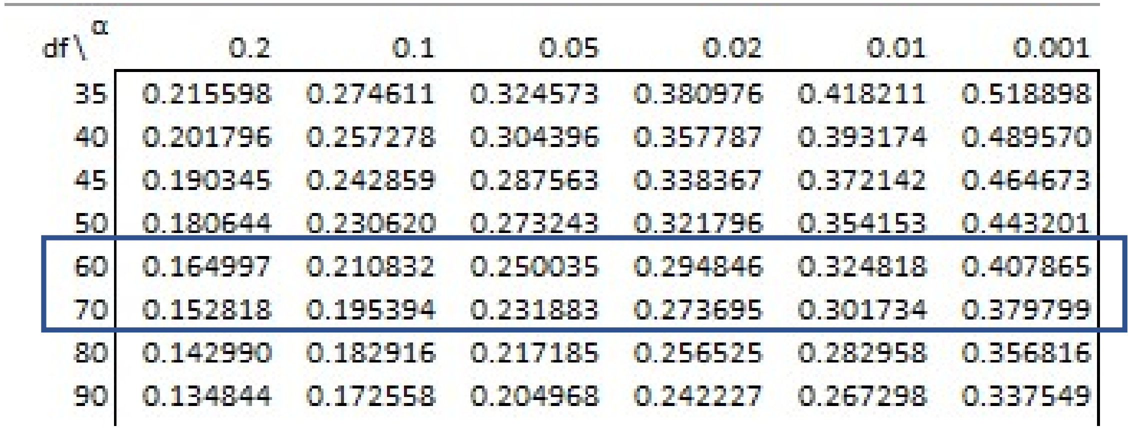
Normal Probability

□ Normal Probability Plots

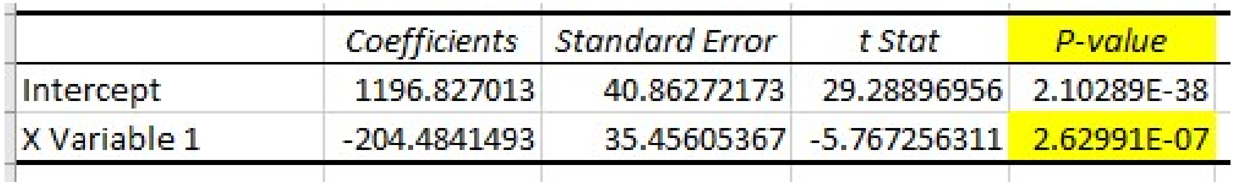
Select 95% Confidence Level

Variable **X** = November- December NINO3.4
Variable **Y** = November-March rainfall
No Blank Cells. Excel returns an error if there are blanks

Using excel for correlation significance



df\α	0.2	0.1	0.05	0.02	0.01	0.001
35	0.215598	0.274611	0.324573	0.380976	0.418211	0.518898
40	0.201796	0.257278	0.304396	0.357787	0.393174	0.489570
45	0.190345	0.242859	0.287563	0.338367	0.372142	0.464673
50	0.180644	0.230620	0.273243	0.321796	0.354153	0.443201
60	0.164997	0.210832	0.250035	0.294846	0.324818	0.407865
70	0.152818	0.195394	0.231883	0.273695	0.301734	0.379799
80	0.142990	0.182916	0.217185	0.256525	0.282958	0.356816
90	0.134844	0.172558	0.204968	0.242227	0.267298	0.337549



	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	1196.827013	40.86272173	29.28896956	2.10289E-38
X Variable 1	-204.4841493	35.45605367	-5.767256311	2.62991E-07

Check for Significance

Our Honiara-NINO3.4 data has **64 degrees of freedom**

Our critical **r** value will lie roughly half way between the two lines highlighted from the table

Our **r** value of –0.58782 is **highly significant**, even at the 0.001 (0.1%level).

Using the Excel Regression Analysis returns a **P value** of **2.63 x 10^{-7}**. This confirms the highly significant nature of the correlation.

Climate and Oceans Support Program in the Pacific

Root Mean Square Error

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are predicted values

$y_1, y_2, \ldots, y_n$ are observed values

n is the number of observations

- RMS stands for **Root Mean Squared**
- It looks similar to the Standard Deviation
- Measures the error of a **model** in **predicting** data
- n = Sample size
- $\hat{y}_i - y_i$ is the **error** (anomaly or residual) between the model prediction and the observation
- Each error is **squared**
- We calculate the **sum** (∑) of all the squared errors
- This sum is divided by the number of observations to create the **mean** of the squared errors
- Finally, calculate the square **root** of the mean

- This is a common method used in ACCESS-S model verification

Significance and Correlation summary

Statistical significance is important for:

1. **Trends** in which **one variable** is plotted against **time** (e.g. climate change)

2. **Correlation** in which **two variables** are plotted against each other (e.g. NINO3.4 and rainfall)

Significance for **trends** is calculated using a linear line of best fit through the data to see how it changes over time (e.g. temperature over time)

Significance for **correlation** measures the strength of a relationship between a dependent and independent variable (e.g. rainfall and NINO3.4)

Excel can be used to calculate statistical significance