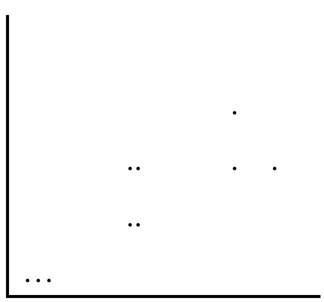
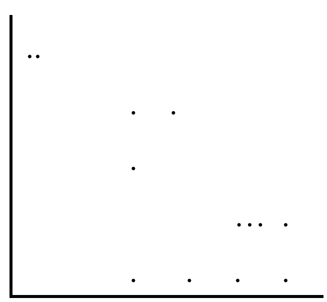
## SIMPLE LINEAR CORRELATION

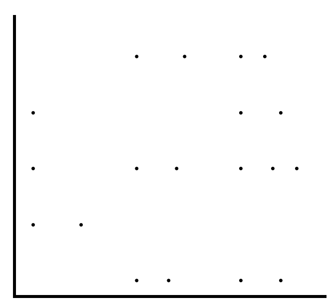
- Simple linear correlation is a measure of the degree to which two variables vary together, or a measure of the intensity of the association between two variables.
- Correlation often is abused. You need to show that one variable actually is affecting another variable.
- The parameter being measure is $\lambda$ (rho) and is estimated by the statistic r, the correlation coefficient.
- r can range from -1 to 1, and is independent of units of measurement.
- The strength of the association increases as r approaches the absolute value of 1.0
- A value of 0 indicates there is no association between the two variables tested.
- A better estimate of r usually can be obtained by calculating r on treatment means averaged across replicates.
- Correlation does not have to be performed only between independent and dependent variables.
- Correlation can be done on two dependent variables.
- The X and Y in the equation to determine r do not necessarily correspond between a independent and dependent variable, respectively.
- Scatter plots are a useful means of getting a better understanding of your data.



Positive association



Negative association



No association

The formula for r is: $$r = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{\sum\left(X-\overline{X}\right)^2 \sum\left(Y-\overline{Y}\right)^2}} = \frac{SSCP}{\sqrt{(SSX)(SSY)}}$$

**Example 1**

| X | Y | XY |
|---|---|---|
| 41 | 52 | 2132 |
| 73 | 95 | 6935 |
| 67 | 72 | 4824 |
| 37 | 52 | 1924 |
| 58 | 96 | 5568 |
| $\sum X = 276$ | $\sum Y = 367$ | $\sum XY = 21{,}383$ |
| $\sum X^2 = 16{,}232$ | $\sum Y^2 = 28{,}833$ | n = 5 |

Step 1. Calculate SSCP

$$SSCP = 21{,}383 - \frac{(276)(367)}{5} = 1124.6$$

Step 2. Calculate SS X

$$SS\ X = 16{,}232 - \frac{276^2}{5} = 996.8$$

Step 3. Calculate SS Y

$$SS\ Y = 28{,}233 - \frac{367^2}{5} = 1895.2$$

Step 4. Calculate the correlation coefficient r

$$r = \frac{SSCP}{\sqrt{(SSX)(SSY)}} = \frac{1124.6}{\sqrt{(996.8)(1895.2)}} = 0.818$$

**<u>Testing the Hypothesis That an Association Between X and Y Exists</u>**

- To determine if an association between two variables exists as determined using correlation, the following hypotheses are tested:

  $H_o$: ƛ = 0
  $H_A$: ƛ ≠0

- Notice that this correlation is testing to see if *r* is significantly different from zero, i.e., there is an association between the two variables evaluated.

- **You are not testing to determine if there is a "SIGNIFICANT CORRELATION".** This cannot be tested.

- Critical or tabular values of *r* to test the hypothesis $H_o$: ƛ = 0 can be found in tables, in which:
  - The df are equal to n-2
  - The number of independent variables will equal one for all simple linear correlation.

- The tabular *r*-value, $r_{.05,\ 3\ df} = 0.878$

- Because the calculated *r* (.818) is less than the table *r* value (.878), we fail to reject $H_o$: ƛ = 0 at the 95% level of confidence. We can conclude that there is no association between X and Y.

- In this example, it would appear that the association between X and Y is strong because the *r* value is fairly high. Yet, the test of $H_o$: ƛ = 0 indicates that there is not a linear relationship.

**Points to Consider**

1. The tabular *r* values are highly dependent on n, the number of observations.

2. As n increases, the tabular r value decreases.

3. We are more likely to reject $H_o$: ƛ = 0 as n increases.

4. As n approaches 100, the r value to reject $H_o$: ƛ = 0 becomes fairly small. Too many people abuse correlation by not reporting the r value and stating incorrectly that there is a “significant correlation”. **The failure to accept $H_o$: ƛ = 0 says nothing about the strength of the association between the two variables measured.**

5. The correlation coefficient squared equals the coefficient of determination. Yet, you need to be careful if you decide to calculate *r* by taking the square root of the coefficient of determination. You may not have the correct “sign” is there is a negative association between the two variables.

ly of the table for corresponding

| .06 | .07 | .08 | .09 |
|---|---|---|---|
| | .07012 | .08017 | .09024 |
| | .17167 | .18198 | .19234 |
| | .27686 | .28768 | .29857 |
| | .38842 | .40006 | .41180 |
| | .51007 | .52298 | .53606 |
| | .64752 | .66246 | .67767 |
| | .81074 | .82911 | .84795 |
| | 1.02033 | 1.04537 | 1.07143 |
| | 1.33308 | 1.37577 | 1.42192 |
| | 2.09229 | 2.29756 | 2.64665 |

Four-figure Mathematical Tables, ...mission of the authors and the

7 8

082



**Table A.13 Significant values of $r$ and $R$**

| Error df | $P$ | Independent variables 1 | 2 | 3 | 4 | Error df | $P$ | Independent variables 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | .05 | .997 | .999 | .999 | .999 | 24 | .05 | .388 | .470 | .523 | .562 |
| | .01 | 1.000 | 1.000 | 1.000 | 1.000 | | .01 | .496 | .565 | .609 | .642 |
| 2 | .05 | .950 | .975 | .983 | .987 | 25 | .05 | .381 | .462 | .514 | .553 |
| | .01 | .990 | .995 | .997 | .998 | | .01 | .487 | .555 | .600 | .633 |
| 3 | .05 | .878 | .930 | .950 | .961 | 26 | .05 | .374 | .454 | .506 | .545 |
| | .01 | .959 | .976 | .983 | .987 | | .01 | .478 | .546 | .590 | .624 |
| 4 | .05 | .811 | .881 | .912 | .930 | 27 | .05 | .367 | .446 | .498 | .536 |
| | .01 | .917 | .949 | .962 | .970 | | .01 | .470 | .538 | .582 | .615 |
| 5 | .05 | .754 | .836 | .874 | .898 | 28 | .05 | .361 | .439 | .490 | .529 |
| | .01 | .874 | .917 | .937 | .949 | | .01 | .463 | .530 | .573 | .606 |
| 6 | .05 | .707 | .795 | .839 | .867 | 29 | .05 | .355 | .432 | .482 | .521 |
| | .01 | .834 | .886 | .911 | .927 | | .01 | .456 | .522 | .565 | .598 |
| 7 | .05 | .666 | .758 | .807 | .838 | 30 | .05 | .349 | .426 | .476 | .514 |
| | .01 | .798 | .855 | .885 | .904 | | .01 | .449 | .514 | .558 | .591 |
| 8 | .05 | .632 | .726 | .777 | .811 | 35 | .05 | .325 | .397 | .445 | .482 |
| | .01 | .765 | .827 | .860 | .882 | | .01 | .418 | .481 | .523 | .556 |
| 9 | .05 | .602 | .697 | .750 | .786 | 40 | .05 | .304 | .373 | .419 | .455 |
| | .01 | .735 | 800 | .836 | .861 | | .01 | .393 | .454 | .494 | .526 |
| 10 | .05 | .576 | .671 | .726 | .763 | 45 | .05 | .288 | .353 | .397 | .432 |
| | .01 | .708 | .776 | .814 | .840 | | .01 | .372 | .430 | .470 | .501 |
| 11 | .05 | .553 | .648 | .703 | .741 | 50 | .05 | .273 | .336 | .379 | .412 |
| | .01 | .684 | .753 | .793 | .821 | | .01 | .354 | .410 | .449 | .479 |
| 12 | .05 | .532 | .627 | .683 | .722 | 60 | .05 | .250 | .308 | .348 | .380 |
| | .01 | .661 | .732 | .773 | .802 | | .01 | .325 | .377 | .414 | .442 |
| 13 | .05 | .514 | .608 | .664 | .703 | 70 | .05 | .232 | .286 | .324 | .354 |
| | .01 | .641 | .712 | .755 | .785 | | .01 | .302 | .351 | .386 | .413 |
| 14 | .05 | .497 | .590 | .646 | .686 | 80 | .05 | .217 | .269 | .304 | .332 |
| | .01 | .623 | .694 | .737 | .768 | | .01 | .283 | .330 | .362 | .389 |
| 15 | .05 | .482 | .574 | .630 | .670 | 90 | .05 | .205 | .254 | .288 | .315 |
| | .01 | .606 | .677 | .721 | .752 | | .01 | .267 | .312 | .343 | .368 |
| 16 | .05 | .468 | .559 | .615 | .655 | 100 | .05 | .195 | .241 | .274 | .300 |
| | .01 | .590 | .662 | .706 | .738 | | .01 | .254 | .297 | .327 | .351 |
| 17 | .05 | .456 | .545 | .601 | .641 | 125 | .05 | .174 | .216 | .246 | .269 |
| | .01 | .575 | .647 | .691 | .724 | | .01 | .228 | .266 | .294 | .316 |
| 18 | .05 | .444 | .532 | .587 | .628 | 150 | .05 | .159 | .198 | .225 | .247 |
| | .01 | .561 | .633 | .678 | .710 | | .01 | .208 | .244 | .270 | .290 |
| 19 | .05 | .433 | .520 | .575 | .615 | 200 | .05 | .138 | .172 | .196 | .215 |
| | .01 | .549 | .620 | .665 | .698 | | .01 | .181 | .212 | .234 | .253 |
| 20 | .05 | .423 | .509 | .563 | .604 | 300 | .05 | .113 | .141 | .160 | .176 |
| | .01 | .537 | .608 | .652 | .685 | | .01 | .148 | .174 | .192 | .208 |
| 21 | .05 | .413 | .498 | .522 | .592 | 400 | .05 | .098 | .122 | .139 | .153 |
| | .01 | .526 | .596 | .641 | .674 | | .01 | .128 | .151 | .167 | .180 |
| 22 | .05 | .404 | .488 | .542 | .582 | 500 | .05 | .088 | .109 | .124 | .137 |
| | .01 | .515 | .585 | .630 | .663 | | .01 | .115 | .135 | .150 | .162 |
| 23 | .05 | .396 | .479 | .532 | .572 | 1,000 | .05 | .062 | .077 | .088 | .097 |
| | .01 | .505 | .574 | .619 | .652 | | .01 | .081 | .096 | .106 | .115 |

SOURCE: Reproduced from G. W. Snedecor, *Statistical Methods*, 4th ed, The Iowa State College Press, Ames, Iowa, 1946, with permission of the author and publisher.

**Example 2**

Assume X is the independent variable and Y is the dependent variable, $n$ = 150, and the correlation between the two variables is $r$ = 0.30. This value of r is significantly different from zero at the 99% level of confidence.

Calculating $r^2$ using r, $0.30^2 = 0.09$, we find that 9% of the variation in Y can be explained by having X in the model. This indicates that even though the $r$ value is significantly different from zero, the association between X and Y is weak.

Some people feel the coefficient of determination needs to be greater that 0.50 (i.e. $r$ = 0.71) before the relationship between X an Y is very meaningful.

**Calculating *r* Combined Across Experiments, Locations, Runs, etc.**

This is another area where correlation is abused.

When calculating the "pooled" correlation across experiments, you **cannot** just put the data into one data set and calculate $r$ directly. The value of r that will be calculated is not a reliable estimate of $\lambda$.

A better method of estimating $\lambda$ would be to:

1. Calculate a value of $r$ for each environment, and
2. Average the $r$ values across environments.

The proper method of calculating a pooled $r$ value is to test the homogeneity of the correlation coefficients from the different locations. If the $r$ values are homogenous, a pooled $r$ value can be calculated.

Example

The correlation between grain yield and kernel plumpness was 0.43 at Langdon, ND; 0.32 at Prosper, ND; and 0.27 at Carrington, ND. There were 25 cultivars evaluated at each location.

Step 1. Make and complete the following table

| Location | n | $r_i$ | $Z'_i$ | $Z'_i - Z'_w$ | $(n_i-3)(Z'_i - Z'_w)^2$ |
|---|---|---|---|---|---|
| Langdon, ND | 25 | 0.43 | 0.460 | 0.104 | 0.238 |
| Prosper, ND | 25 | 0.32 | 0.332 | -0.024 | 0.013 |
| Carrington, ND | 25 | 0.27 | 0.277 | -0.079 | 0.137 |
| | $\sum n_i=75$ | | $Z'_w = 0.356$ | | $\chi^2 = 0.388$ |

Where:

$$Z_i' = 0.5\ln\left[\frac{(1+r_i)}{(1-r_i)}\right]$$

$$Z_w' = \frac{\sum[(n_i-3)Z_i']}{\sum(n_i-3)}$$

$$\chi^2 = \sum[(n_i-3)(Z_i'-Z_w')^2]$$

$$\mathrm{df} = n-1 \text{ for } \chi^2 \text{ test}$$

Step 2. Look up tabular $\chi^2$ value at the $\alpha = 0.005$ level.

$\chi^2_{0.005,\ 2\ df} = 10.6$

Step 3. Make conclusions

Because the calculated $\chi^2$ (0.388) is less than the table $\chi^2$ value (10.6), we fail to reject the null hypothesis that the *r*-values from the three locations are equal.

Step 4. Calculate pooled $r$ ($r_p$) value

$$r_p = \frac{e^{2Z_w^{i}} - 1}{e^{2Z_w^{'}} + 1}$$

Where e = 2.71828128

$$\text{Therefore } r_p = \frac{e^{2(0.356)} - 1}{e^{2(0.356)} + 1} = 0.341$$

Step 5. Determine if $r_p$ is significantly different from zero using a confidence interval.

$$r_p \pm 1.96\left(\frac{1}{\sqrt{\sum(n_i - 3)}}\right)$$

$$\text{CI} = \quad = 0.341 \pm 1.96\frac{1}{\sqrt{66}}$$

$$= 0.341 \pm 0.241$$

Therefore LCI = 0.100 and UCI = 0.582

Since the CI does not include zero, we reject the hypothesis that the pooled correlation value is equal to zero.

# SAS Commands for Simple Linear Correlation

```
options pageno=1;
data corr;
input x y;
datalines;
41 52
73 95
67 72
37 52
58 96
;;
ods rtf file ='example.rtf';
run;
proc corr;
var y x;
*Comment: This analysis will provide you with the correlation
coefficient and a test of the null hypothesis that there is no linear
relationship between
the two variable';
title 'Simple Linear Correlation Analysis';
run;
ods rtf close;
run;
```

## *Simple Linear Correlation Analysis*

### *The CORR Procedure*

| **2 Variables:** | y<br>x |
|---|---|

| **Simple Statistics** | | | | | | |
|---|---|---|---|---|---|---|
| **Variable** | **N** | **Mean** | **Std Dev** | **Sum** | **Minimum** | **Maximum** |
| **y** | 5 | 73.40000 | 21.76695 | 367.00000 | 52.00000 | 96.00000 |
| **x** | 5 | 55.20000 | 15.78607 | 276.00000 | 37.00000 | 73.00000 |

| **Pearson Correlation Coefficients, N = 5<br>Prob > \|r\| under H0: Rho=0** | | |
|---|---|---|
| | **y** | **x** |
| **y** | 1.00000 | 0.81821<br>0.0905 |
| **x** | 0.81821<br>0.0905 | 1.00000 |

- The top value is the correlation value and the bottom value is the Prob>|$r$| to test the null hypothesis $H_o$: $\rho$=0.
- The values on the diagonal are always 1.0.
- The values above and below the diagonal are symmetrical.