Title

Correlation

Overview

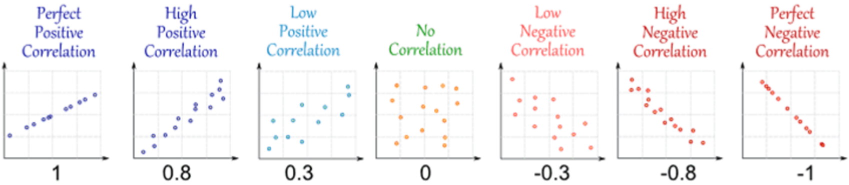
- Correlation is made of **Co-** (meaning ”together”), and **Relation**
- Statistical procedure used to measure and describe the relationship between two variables
- Range between +1 and -1
  - Positive when the values increase together
  - Negative when one value decreases as the other increases

. . .

Overview cont..

- +1 is a perfect positive correlation
- 0 is no correlation (independence)
- -1 is a perfect negative correlation

Like this:



Use of Corelation

When two variables, let's call them X Y, are correlated, then one variable can be used to predict the other variable
Example:IQ and perfomance...

Types

- **Pearson product-moment correlation** -When both variables, X and Y, are continuous
- **Point bi-serial correlation** - When 1 variable is continuous and 1 is dichotomous
- **Phi coefficient** - When both variables are dichotomous
- **Spearman rank correlation** - When both variables are ordinal (ranked data)

Calculation of Correlation

defined as

$$r = S_{xy} / \sqrt{S_{xx} S_{yy}}.$$

where

$$S_{xx} = \sum_{i=1}^{N} (x_i - \bar{x})^2 \text{ (variance of x)}$$

and

$$S_{xy} = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \text{ (covariance of x and y)}$$

Output

```
print(df)


##    temp icecream
## 1  14.2      215
## 2  16.4      325
## 3  11.9      185
## 4  15.2      332
## 5  18.5      406
## 6  22.1      522
## 7  19.4      412
## 8  25.1      614
## 9  23.4      544
## 10 18.1      421
## 11 22.6      445
## 12 17.2      408
```

Output

```
print(df)


##    temp icecream deviationTemp deviationIce      SSxy     SSxx     SSyy
## 1  14.2      215        -4.475     -187.417  838.6896 20.02563 35125.01
## 2  16.4      325        -2.275      -77.417  176.1229  5.17563  5993.34
## 3  11.9      185        -6.775     -217.417 1472.9979 45.90063 47270.01
## 4  15.2      332        -3.475      -70.417  244.6979 12.07563  4958.51
## 5  18.5      406        -0.175        3.583   -0.6271  0.03063    12.84
## 6  22.1      522         3.425      119.583  409.5729 11.73063 14300.17
## 7  19.4      412         0.725        9.583    6.9479  0.52562    91.84
## 8  25.1      614         6.425      211.583 1359.4229 41.28063 44767.51
## 9  23.4      544         4.725      141.583  668.9812 22.32562 20045.84
## 10 18.1      421        -0.575       18.583  -10.6854  0.33062   345.34
## 11 22.6      445         3.925       42.583  167.1396 15.40563  1813.34
## 12 17.2      408        -1.475        5.583   -8.2354  2.17563    31.17


print(sum.SSxy)


## [1] 5325


print(sum.SSxx)
```

Corelation in R

```
## [1] 177
```

```r
print(sum.SSyy)
```

```
## [1] 174755
```

Corelation in R

```
cor(df$temp, df$icecream)
```

```
## [1] 0.9575
```

```
cor.test(df$temp, df$icecream)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$temp and df$icecream
## t = 10.5, df = 10, p-value = 1.016e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8515 0.9883
## sample estimates:
##    cor
## 0.9575
```

Diff btwn cor and cor.test The cor.test output also includes the point estimate reported by cor
Cor.test has p-value and also CI

Caution

- **!”Correlation Is Not Causation” ...**
  When there is a correlation it does not mean that one thing causes the other
- The magnitude of a correlation depends upon many factors, including
  - Sampling (random and representative?)
  - Measurement of X and Y and Several other assumptions . . .

. . .

Assumptions

- Normal Distribution for X and Y if not specifying the method - Use method="Spearman" for non-normal data.
- Linear relationship between X and Y
- **Homoscedasticity** - homogeneity of variance/ uniformity of variance leveneTest() from car package is used to test this