

Equations!

Deviation score:	$(x - \bar{x})$
Squared deviation score:	$(x - \bar{x})^2$
Sum of squares:	$SS = \sum (x - \bar{x})^2$
Variance:	$SD^2 = SS \div N$
Standard deviation:	$\sqrt{\text{variance or } SD^2}$
Covariance	$cov = SP \div N$
Pearson correlation:	$r = cov \div (SD_x)(SD_y)$
Slope:	$b_y = r(SD_y \div SD_x)$
intercept:	$a_y = \bar{y} - b_y(\bar{x})$
Total variability:	$SST = \sum (Y - \bar{y})^2$
explained variability:	$SSR = \sum (Y' - \bar{y})^2$
unexplained variability	$SSE = \sum (Y - Y')^2$
Standard error of prediction:	$SD_{y-y'} = SD_y \sqrt{1 - r^2}$
Predicting X':	$X' = a_x + b_x Y$
Predicting Y':	$Y' = a_y + b_y X$

General guidelines for test reliability

>.85	very desirable
.70 to .85	desirable aka moderately acceptable
<.70	not desirable aka poor reliability

describe relationship between two variables?

1.) Direction of the relationship:

Positive (+) or negative (-)

Positive correlation = As the values of x increase or decrease, so do the values of y
No relationship = no consistent relationship between variables

Negative correlation = As the values of x increases, the value of y decreases, and vice versa

2.) shape of the relationship

describe relationship between two variables? (cont)

Linear relationship = straight line relationships
– All dots clustered around straight line
Curvilinear relationship = consistent, predictable relationship, but not linear
– As the values of x increase, the values of y increases but at some point the pattern reverses

3.) Strength of the relationship

Subjective measure of relationship between two scores (e.g., weak, moderate, strong, no relationship)
how closely the data points cluster together
The more spread out they are from a line of some sort, the weaker the correlation between variables

4.) Magnitude of the relationship

Objective measure of relationship based on computed r value: ranges from -1 to 1

biserial correlation

When to use it:

– when one of the variables is nominal (with only two groups) and the other variable is interval/ratio

How to calculate:

– use the same formula as pearson r

Curvilinear relationships:

Linear: $Y' = a + bX$

Quadratic: $Y' = a + bX + cX^2$

Cubic: $Y' = a + bX + cX^2 + dX^3$

Quartic: $Y' = a + bX + cX^2 + dX^3 + eX^4$

Comparing SD_{y-y'} and SD_y

When R does not equal Zero, SD_{y-y'} will be smaller than SD_y

When R=0 (no correlation/relationship), SD_{y-y'} = SD_y

When R=+/- 1 (perfect correlation), SD_{y-y'}=0

How do we describe our data?

- 1.) plotting a scatter plot, linearity, Shape strength, direction, magnitude
- 2.) defining the regression line Central (mean of bivariate data) tendency
- 3.) standard error or estimates Variability (SD_{y-y'})

Factors affecting R

- 1.) Relationship is real and strong or weak contributes to a bigger/smaller r
- 2.) Sampling error Sampling error = naturally occurring discrepancy, or error, that exists between a sample statistic and the corresponding parameter
- 3.) Unmeasured third variable contributes to a bigger/smaller r, Correlation tells us if a relationship between two variables exists but does not tell us about causation
- 4.) Heterogeneous sample Data in which the sample of observations could be subdivided into two distinct sets on the basis of some other variable
- 5.) Sampling from a restricted (truncated) range The correlation coefficient will be affected by the range of score in the data

Factors affecting R (cont)

- 6.) Non-linearity: relationship is curvilinear
Reminder: r underestimates a curvilinear relationship, contributes to a smaller r
- 7.) Heteroscedasticity in the data
contributes to a smaller r

PHI

When to use it:

- when both variables are nominal (with only two groups per variable, i.e., dichotomous)

Calculating Phi:

- use the same formula as Pearson r

How to calculate Pearson r:

- 1.) Plot the data (scatterplot)
- 2.) Compute (e.g., deviation scores, bivariate statistics SP, COV)
- 3.) Compute (number beyond +/-1 correlation coefficient r means you did it wrong)

Interpreting Pearson Correlation

- | | |
|-----------------|-----------------------|
| < .10 | no relationship |
| .10 to .30 | weak relationship |
| > .30 to .50 | moderate relationship |
| > .50 | strong relationship |

Reporting in APA format

- 1.) Give variables, $R = ?$, Mean = ?, Standard deviation = ?, Give relationship in statistical terms and if its positive for negative
- 2.) Results in plain language

extra stuff

Homoscedasticity (a good thing):
Variability in Y scores remains constant across increasing values of X

Heteroscedasticity (not a good thing):
variability in y scores changes across increasing values of x ,
Caused by a skew in one or both variables

$$SST = SSy$$

$$SSE = SSy - y'$$

(error)

$$SSr = SST - SSE$$

$$\Sigma(Y - Y') = 0$$

For Y' : if $r=0$, $b_y=0$ (i.e., regression line is parallel to the x -axis), and $ay=\bar{y}$

For X' : if $r=0$, $b_x=0$ (i.e., regression line is parallel to the x -axis), and $ax=\bar{x}$

As correlation (r) increases, the numerical value for b increases

Total variability = differences between observed data (Y) and the mean value of Y

Unexplained variability (i.e., residuals) = difference between the observed value for Y and the predicted value for $Y(Y')$

Explained variability = the difference between total and unexplained variability

Standardized test = interval

Spearman rho

When to use it:

- one or both variables are on an ordinal scale of measurement
- there is a weak curvilinear relationship in interval/ratio data
- there is heteroscedasticity in interval/ratio data

How to calculate:

Convert all scores into ranks

Lower scores get lower ranks

High scores get higher ranks

Use the Pearson correlation formula to find how consistently increases in one variable are associated with increases in another variable