CORRELATION & REGRESSION

Definition of Correlation:

Relationship between variables:

Examples:

Characteristics:

- *Nonmanipulative*
- *Pairs of scores*
- *Presentation: Tables or Scatterplots*

CORRELATIONS MEASURE 3 CHARACTERISTICS OF
THE RELATIONSHIP BETWEEN X & Y:

1) The Direction of the Relationship

Positive correlation:

Negative correlation/inverse relationship:

ACTIVITY #1: POSITIVE AND NEGATIVE CORRELATIONS

Directions: *Tell whether each of the following relationships is likely to be positive or negative.*

1) The amount of stress in people's lives and the number of colds they get in the winter
2) The amount of time people spend sun-tanning and a dermatological index of skin damage
3) Happiness and suicidal thoughts
4) Blood pressure and level of hostility
5) The number of times a rat has run through a maze and the time it takes to run it again

2) The Form of the Relationship

No Relationship:

Linear Relationship:

Curvilinear Relationship:

Monotonic Relationship:

Nonmonotonic Relationship:

Caveat:

3) The Degree of the Relationship

Correlation Coefficients:

- *Definition*
- *Direction of relationship:*
- *Magnitude/Strength of Relationship:*

ACTIVITY #2: DESCRIBING CORRELATION COEFFICIENTS

Directions: *Describe in words the nature of the relationship between each pair of variables as indicated by the value of the correlation coefficient. Be sure to include the direction of the relationship, the strength of the relationship, and a verbal description of the way the variables "go together."*

1) $r = -0.96$ between craving for pizza and ability to concentrate on studying
2) $r = +0.02$ between length of marriage and marital satisfaction
3) $r = +0.55$ between parent and child intelligence test scores
4) $r = -0.90$ between amount of alcohol consumed and performance on a motor coordination task
5) $r = +0.75$ between amount of money won and the number of hot dogs purchased at the race tracks

6) r = +0.67 between scores on a hyperactivity scale and scores on an aggressiveness scale

7) r = -0.82 between a job applicant's age and likelihood of being hired

8) r = +0.06 between the number of yearly predictions by psychics and the number of correct predictions

THE PEARSON PRODUCT MOMENT CORRELATION

Introduction:

$$r = \frac{\text{degree to which X \& Y vary together (covariability of X \& Y)}}{\text{degree to which X \& Y vary separately (variability of X \& Y)}}$$

The Sum of Products of Deviations:

Definitional formula: SP = $\sum \left(X - \overline{X}\right)\left(Y - \overline{Y}\right)$

Computational formula: SP = $\sum XY - \frac{\sum X \sum Y}{n}$

Example:

Scores

X	Y
1	3
2	6
4	4
5	7

Calculating the Pearson Correlation:

$$r = \frac{SP}{\sqrt{SS_X SS_Y}}$$

Example:

Scores	
X	Y
0	1
10	3
4	1
8	2
8	3
ΣX=	ΣY=
$\overline{X}$ =	$\overline{Y}$ =
ΣX^2=	ΣY^2=
ΣXY=	

$$SS_X = \sum X^2 - \frac{\left(\sum X\right)^2}{n} =$$

$$SS_Y = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n} =$$

$$SP = \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n} =$$

Graph scores first and then calculate *r*

$$r = \frac{SP}{\sqrt{SS_X SS_Y}} =$$

Using & Interpreting the Pearson Correlation

Where & Why Correlations Are Used:

1. *Prediction*

2. *Validity*

3. *Reliability*

4. *Theory Verification*

Interpreting Correlations:

1. Correlation & Causation

- *Direction of cause and effect:*

- *The third variable problem:*

2. Correlation & Restricted Range

3. Outliers

4. Correlation & the Strength of the Relationship

The coefficient of determination

<u>Hypothesis Tests with the Pearson Correlation</u>

The Hypotheses:

H_0: $\rho = 0$

H_1: $\rho \neq 0$

The purpose of the hypothesis test is to decide between the following 2 alternatives:

1)

2)

Degrees of Freedom:

$df = n - 2$

ACTIVITY #3: PRACTICE WITH STATISTICAL SIGNIFICANCE

1. Using Table E (Critical Values of *r*), indicate whether each of the following correlation coefficients is statistically significant.

   $r = .42$, $n = 112$

   $r = .00$, $n = 1000$

   $r = -.25$, $n = 50$

   $r = -.15$, $n = 100$

   $r = .05$, $n = 300$

   $r = .25$, $n = 60$

2. You administered measures of anxiety and insomnia to a sample of 30 participants and obtained a correlation coefficient of .28.

   A. Is this correlation statistically significant?

   B. What is the critical value?

   C. If the correlation is not significant, but you still think that anxiety and insomnia are correlated, what could you do to conduct the study again, providing a more powerful test of your hypothesis?

MEASURING THE RELATIONSHIP BETWEEN TWO VARIABLES

1. Measure the variables of interest.

2. Plot the data on a scatterplot.

3. In order to quantify the degree of relationship more precisely, calculate a correlation coefficient.

4. Set up a table for recording data and carrying out computations.

Subject #	**X**	**Y**	**XY**	**X^2**	**Y^2**
S_1					
S_2					
S_3					
…					
S_N					
Totals	$\sum X$	$\sum Y$	$\sum XY$	$\sum X^2$	$\sum Y^2$

5. Fill in the table.

Example: The Relationship Btn SAT Quantitative Scores & Final Exam Scores in Intro to Psych

Subject #	**X**	**Y**	**XY**	**X^2**	**Y^2**
1	595	68	40,460	354,025	4,624
2	520	55	28,600	270,400	3,025
3	715	65	46,475	511,225	4,225
4	405	42	17,010	164,025	1,764
5	680	64	43,520	462,400	4,096
6	490	45	22,050	240,100	2,025
7	565	56	31,640	319,225	3,136
8	580	59	34,220	336,400	3,481
9	615	56	34,440	378,225	3,136
10	435	42	18,270	189,225	1,764
11	440	38	16,720	193,600	1,444
12	515	50	25,750	265,225	2,500
13	380	37	14,060	144,400	1,369
14	510	42	21,420	260,100	1,764
15	565	53	29,945	319,225	2,809
Totals	$\sum X =$ 8010	$\sum Y =$ 772	$\sum XY =$ 424,580	$\sum X^2 =$ 4,407,800	$\sum Y^2 =$ 41,162

6. Compute the totals for each column.

7. Substitute the totals and the value for N (the number of subjects) into the formula for r.

$$r = \frac{\sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}\right)}}$$

8. Compare the r obtained to the critical value for r.

9. Square r to determine the proportion of variance in Y scores that can be accounted for by knowing X. r^2 is called the coefficient of determination.

THE POINT BISERIAL CORRELATION

Purpose:

Dichotomous variables:

Computing the point-biserial correlation:

Participant	Memory Score		Treatment Condition (Images or No Images)		
	X	X^2	Y	Y^2	XY
A	19		1		
B	20		1		
C	24		1		
D	30		1		
E	31		1		
F	32		1		
G	30		1		
H	27		1		
I	22		1		
J	25		1		
K	23		0		
L	22		0		
M	15		0		
N	16		0		
O	18		0		
P	12		0		
Q	16		0		
R	19		0		
S	14		0		
T	25		0		
	$\sum X =$ n = $\overline{X} =$	$\sum X^2 =$	$\sum Y =$ n = $\overline{Y} =$	$\sum Y^2 =$	$\Sigma XY=$

$$SS_X = \sum X^2 - \frac{\left(\sum X\right)^2}{n} =$$

$$SS_Y = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n} =$$

$$SP = \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n} =$$

$$r = \frac{SP}{\sqrt{SS_X SS_Y}} =$$

r measures:

r-obtained = *r-crit ( ) = , α=.05;* *r-crit ( ) = , α=.01*

$r^2 =$

r^2 measures:

THE SPEARMAN CORRELATION

What It Is Used For:

Example:

Person	X	Y	X rank	Y rank
A	4	9		
B	2	2		
C	10	10		
D	3	8		

Converting raw scores to ranks:

Important Note:

Calculation of the Spearman Correlation (r_s):

1)

2)

3)

Original Data			**Ranks**			
X	Y	X-rank	X^2	Y-rank	Y^2	XY
3	12					
4	10					
8	11					
10	9					
13	3					
		$\Sigma X=$	$\Sigma X^2=$	$\Sigma Y=$	$\Sigma Y^2=$	$\Sigma XY=$

$$ss_x = \sum X^2 - \frac{\left(\sum X\right)^2}{n} \qquad ss_Y = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n}$$

$$SP = \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n} \qquad r_s = \frac{SP}{\sqrt{(SS_X)(SS_Y)}}$$

r-crit:

Interpretation:

Ranking Tied Scores:

Scores	Rank position	Final rank
3		
3		
5		
6		
6		
6		
12		

ACTIVITY #4: PRACTICE WITH CORRELATIONS

1. A common concern for students (and teachers) is the assignment of grades for essay or term papers, as these grades must be based on subjective judgment of quality. To demonstrate that these judgments actually are valid, an English instructor asked a colleague to rank order a set of terms papers. The ranks and the instructor's grades for these papers are as follows:

Rank	Grade
1	A
2	B
3	A
4	B
5	B
6	C
7	D
8	C
9	C
10	D
11	F

a. Calculate the Spearman correlation for these data. (Note: You must convert the letter grades to ranks & re-order the original rankings so that the best paper has the highest rank)

b. Based on this correlation, does it appear that there is reasonable agreement between the two instructors in their judgment of the papers?

2. To test the effectiveness of a new studying strategy, a psychologist randomly divides a sample of 8 students into two groups, with $n = 4$ in each group. The students in one group receive training in the new studying strategy. Then all students are given 30 minutes to study a chapter from a history textbook before they take a quiz on the chapter. The quiz scores for the two groups are as follows:

Training	No Training
9	4
7	7
6	3
10	6

a. Convert the data into a form suitable for the point-biserial correlation. (Use X = 0 for no training, and the quiz score for Y.)

b. Calculate the point-biserial correlation for this data.

INTRODUCTION TO REGRESSION

Pearson Correlation:

Purpose of the Line:

1)

2)

3)

Goal:

Linear Equations:

Example:
A local tennis club charges a fee of $5/hr plus an annual membership fee of $25. What is the total cost of playing tennis?

To Graph the Line:

Regression:

Regression Line:

The Least-Squares Solution:

$$\hat{Y} = \text{bX} + \text{a}$$

$$b = \frac{SP}{SS_X} \qquad a = \overline{Y} - b\overline{X}$$

Example:

X	Y	$X-\overline{X}$	$Y-\overline{Y}$	$(X-\overline{X})(Y-\overline{Y})$	$(X-\overline{X})^2$
7	11				
4	3				
6	5				
3	4				
5	7				0
$\sum X =$ $\overline{X} =$	$\sum Y =$ $\overline{Y} =$			SP=	SS_X=

$$SP = \sum (X-\overline{X})(Y-\overline{Y}) =$$

$$SS_X = \sum (X-\overline{X})^2 =$$

$$b = \frac{SP}{SS_X} =$$

$$a = \overline{Y} - b\overline{X} =$$

Resulting regression line: $\hat{Y}$ =

<u>Common use of regression equations: Prediction</u>

A few cautions when interpreting the predicted value:

1)

2)

STATISTICAL PROCEDURES BASED ON CORRELATIONS

Multiple Regression

Path Analysis

Factor Analysis

Structural Equation Modeling