

Chapter 15: Correlation

So far...

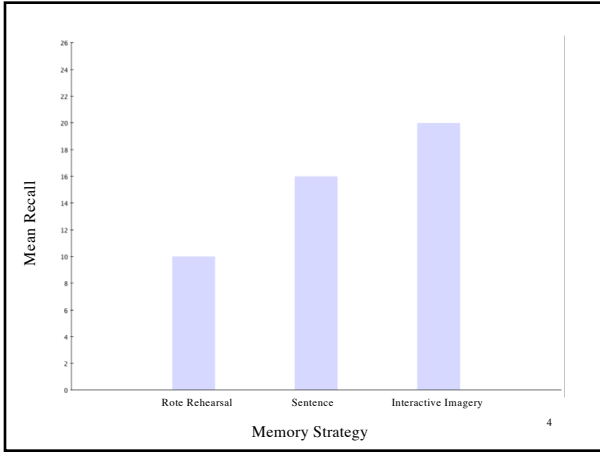
- We' ve focused on hypothesis testing
- Is the *relationship* we observe between x and y in our *sample* true generally (i.e. for the *population* from which the sample came)
- Which answers the following question: *Is there a relationship between x and y?* (*Yes* or *No*)
- Where x is a categorical predictor and y is a continuous predictor

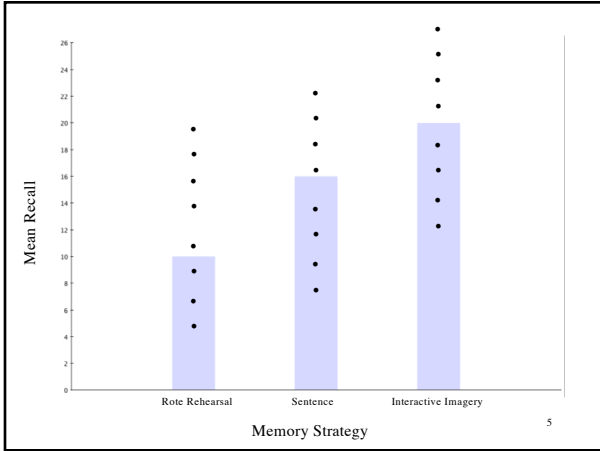
2

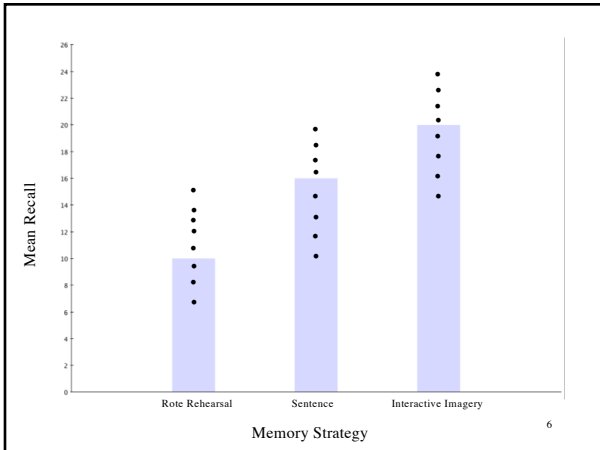
A new question...

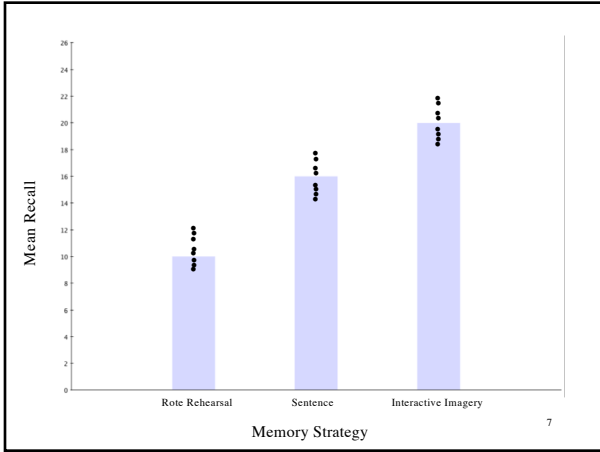
- If there is a relationship between x and y...
- How strong is that relationship?
- How well can we predict a person' s y score if we know x?
- What is the *strength of the relationship* or the *correlation* between x and y

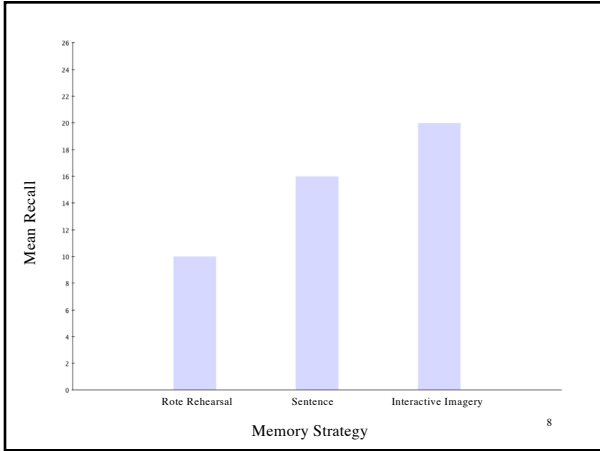
3

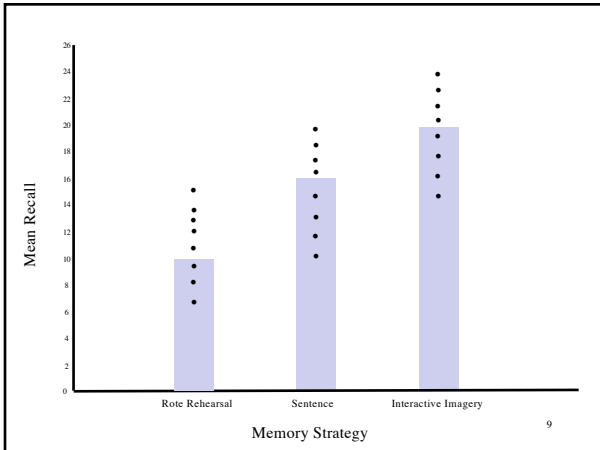


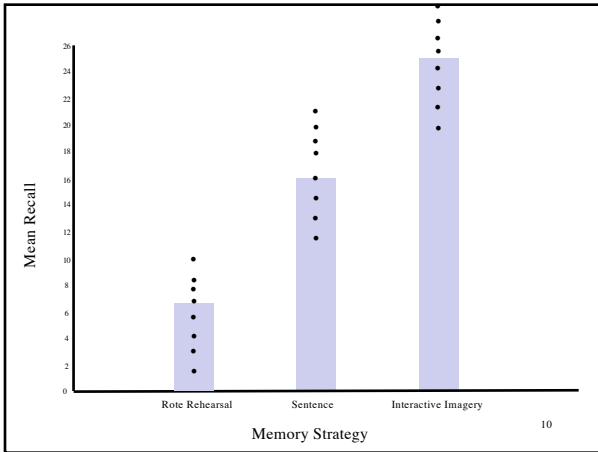






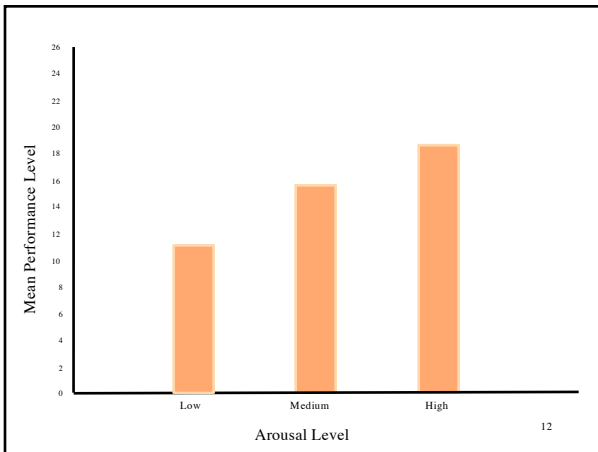


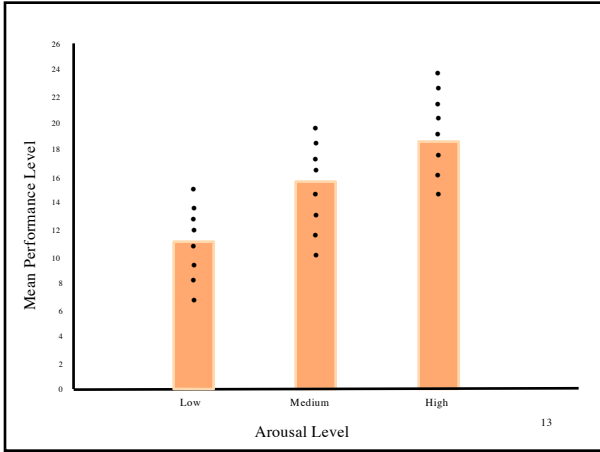


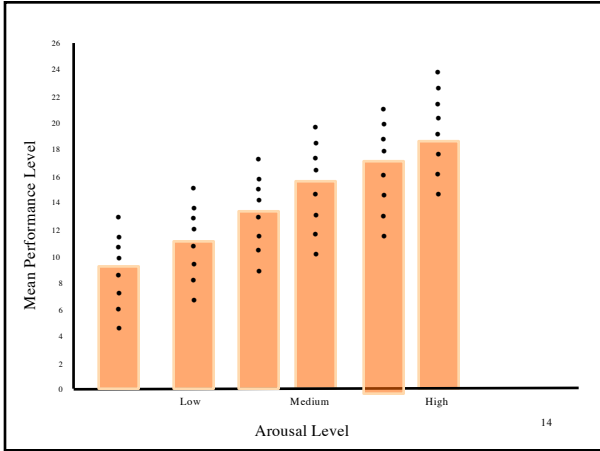


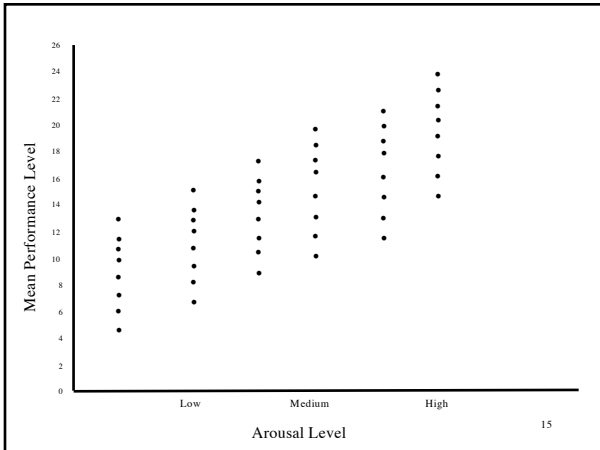
However...

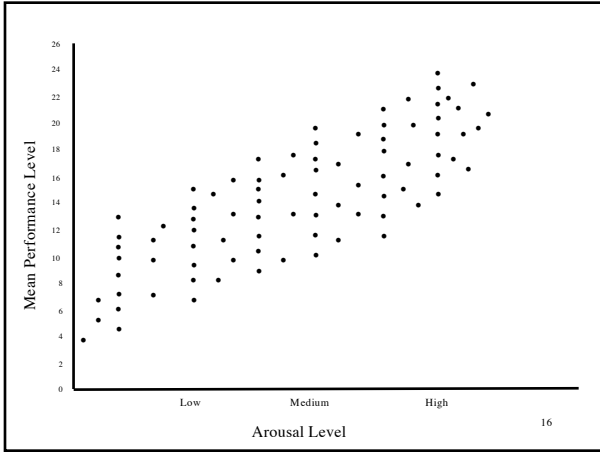
- What if x is not a categorical variable
- What if x is a continuous predictor...e.g. *arousal level*
- And y is a continuous variable as well... e.g. *performance level*

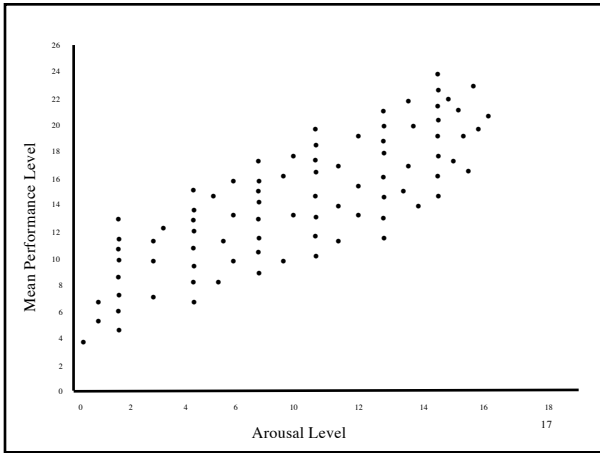


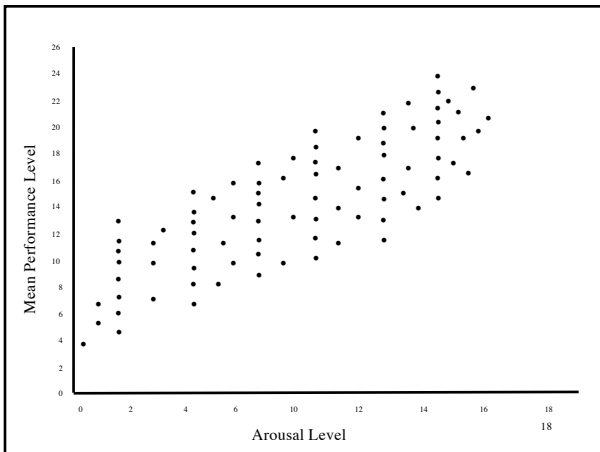


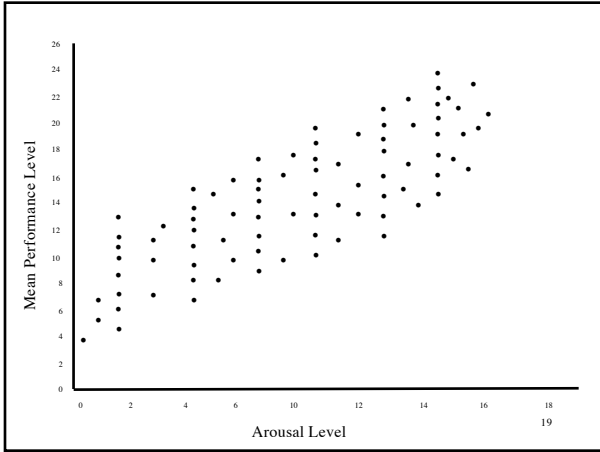


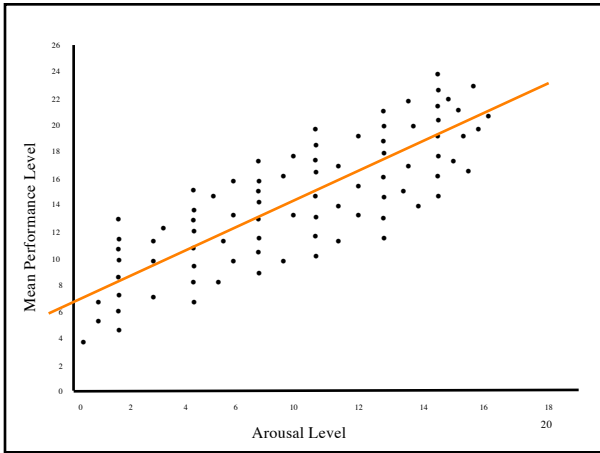


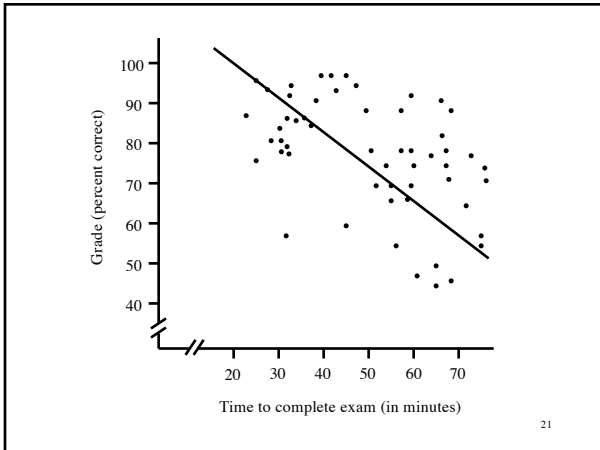


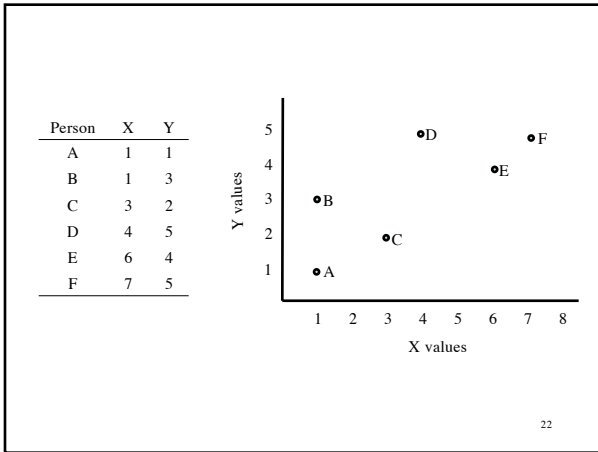












3 Characteristics of a Correlation:

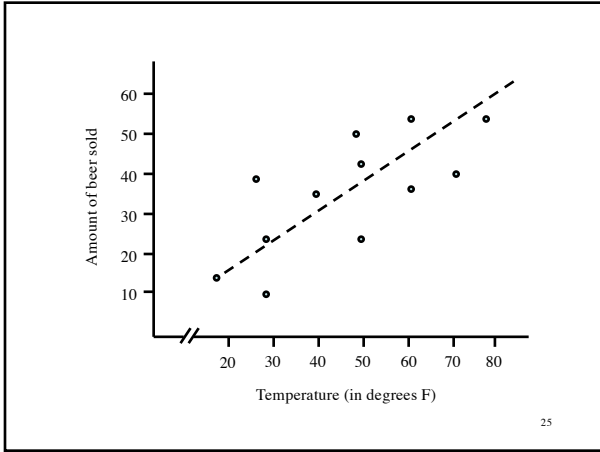
- Direction of relationship
- Form of the relation
- Degree of the relationship

23

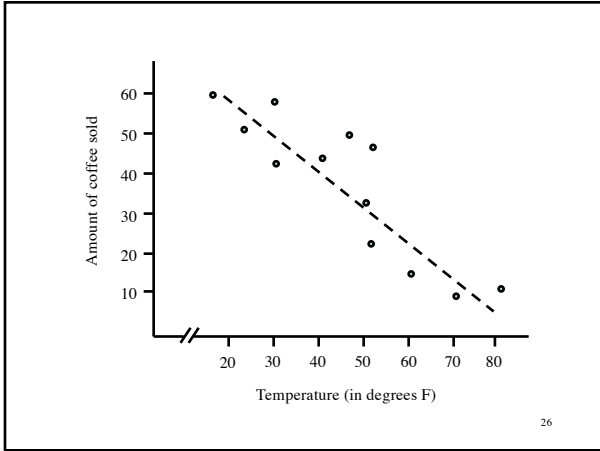
Correlations: Measuring and Describing Relationships (cont.)

- The **direction** of the relationship is measured by the sign of the correlation (+ or -). A positive correlation means that the two variables tend to change in the same direction; as one increases, the other also tends to increase. A negative correlation means that the two variables tend to change in opposite directions; as one increases, the other tends to decrease.

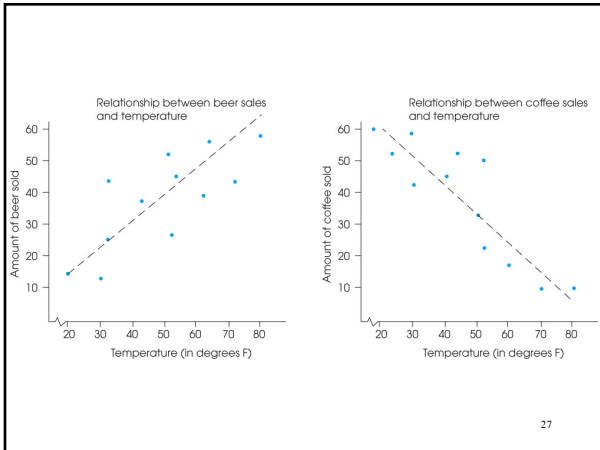
24



25



26

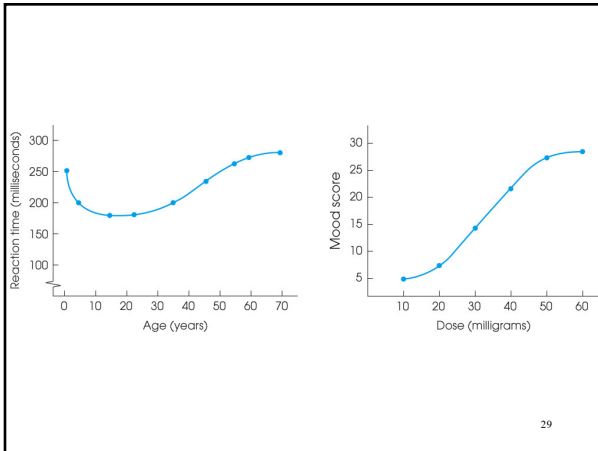


27

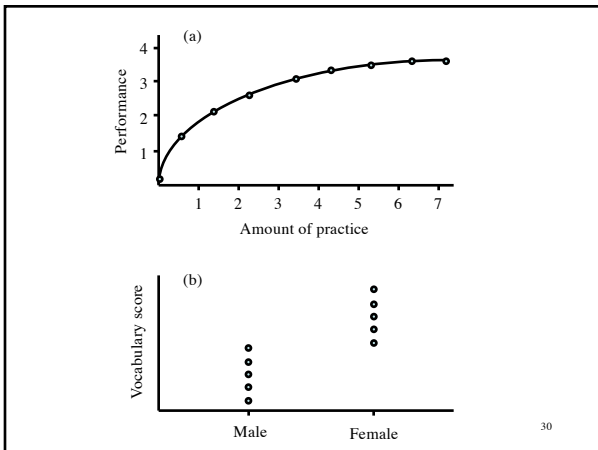
Correlations: Measuring and Describing Relationships (cont.)

- The most common **form** of relationship is a straight line or linear relationship which is measured by the Pearson correlation.

28



29

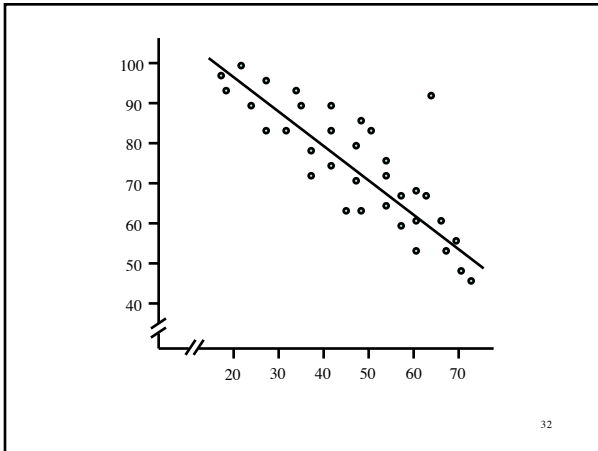


30

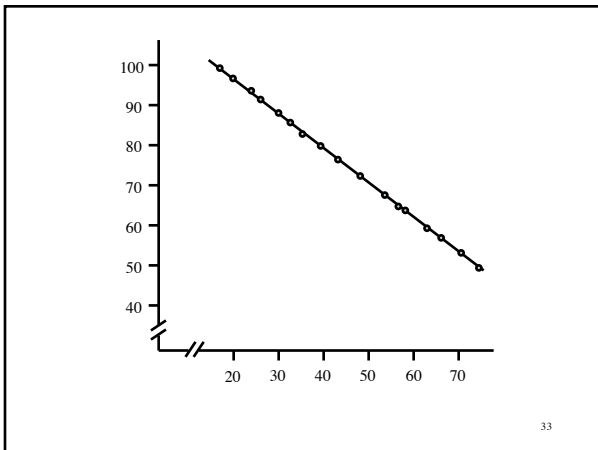
Correlations: Measuring and Describing Relationships (cont.)

- The **degree** of relationship (the strength or consistency of the relationship) is measured by the numerical value of the correlation. A value of 1.00 indicates a perfect relationship and a value of zero indicates no relationship.

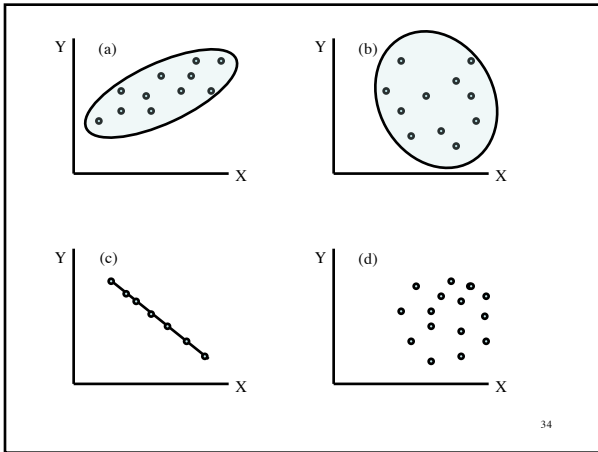
31



32



33



Where and Why Correlations are Used:

- Prediction
- Validity
- Reliability
- Theory Verification

Correlations: Measuring and Describing Relationships (cont.)

- To compute a correlation you need two scores, X and Y, for each individual in the sample.
- The Pearson correlation requires that the scores be numerical values from an interval or ratio scale of measurement.
- Other correlational methods exist for other scales of measurement.

The Pearson Correlation

- The **Pearson correlation** measures the direction and degree of linear (straight line) relationship between two variables.
- To compute the Pearson correlation, you first measure the variability of X and Y scores separately by computing SS for the scores of each variable (SS_x and SS_y).
- Then, the covariability (tendency for X and Y to vary together) is measured by the sum of products (SP).
- The Pearson correlation is found by computing the ratio:

$$r = \frac{SP}{\sqrt{(SS_x)(SS_y)}}$$

37

The Pearson Correlation (cont.)

- Thus the Pearson correlation is comparing the amount of covariability (variation from the relationship between X and Y) to the amount X and Y vary separately.
- The magnitude of the Pearson correlation ranges from 0 (indicating no linear relationship between X and Y) to 1.00 (indicating a perfect straight-line relationship between X and Y).
- The correlation can be either positive or negative depending on the direction of the relationship.

38

The Pearson Correlation

$$r = \frac{\text{degree to which x and y vary together}}{\text{degree to which x and y vary separately}}$$

$$r = \frac{\text{covariability of x and y}}{\text{variability of x and y separately}}$$

$$r = \frac{SP}{\sqrt{(SS_x)(SS_y)}}$$

$$SP = \sum(x - \bar{x})(y - \bar{y})$$

$$SP = \sum xy - \frac{\sum x \sum y}{n}$$

39

The Pearson Correlation

$r = \frac{\text{degree to which } x \text{ and } y \text{ vary together}}{\text{degree to which } x \text{ and } y \text{ vary separately}}$

$r = \frac{\text{covariability of } x \text{ and } y}{\text{variability of } x \text{ and } y \text{ separately}}$

$$r = \frac{SP}{\sqrt{(SS_x)(SS_y)}} \quad SP = \sum (x - \bar{x})(y - \bar{y})$$

$$SP = \sum xy - \frac{\sum x \sum y}{n}$$

PSY version

40

Computational Examples

41

Computing the SP

Scores		Deviations		Products
x	y	(x - \bar{x})	(y - \bar{y})	(x - \bar{x})(y - \bar{y})
1	3			
2	6			
4	4			
5	7			

42

Computing the SP

Scores		Deviations		Products
X	y	(X - \bar{X})	(y - \bar{y})	(X - \bar{X})(y - \bar{y})
1	3	-2	-2	+4
2	6	-1	+1	-1
4	4	+1	-1	-1
5	7	+2	+2	+4
				+6 = SP

43

Computing the SP with the Computational Formula

X	y	xy
1	3	3
2	6	12
4	4	16
5	7	35

$\sum x = 12$ $\sum y = 20$ $\sum xy = 66$

$$SP = \sum xy - \frac{\sum x \sum y}{n}$$

$$SP = 66 - \frac{12(20)}{4}$$

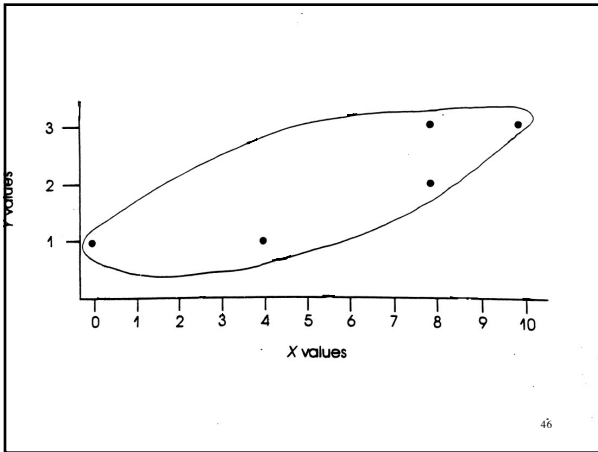
$$SP = 66 - 60$$

$$SP = +6$$

44

X	Y
0	1
10	3
4	1
8	2
8	3

45



Computing a Pearson Correlation

Scores

X	Y
1	3
2	6
4	4
5	7

1. First draw a scatterplot of the x and y data pairs.
2. Then compute the Pearson r correlation coefficient
3. Compare the scatterplot to the calculated Pearson r

47

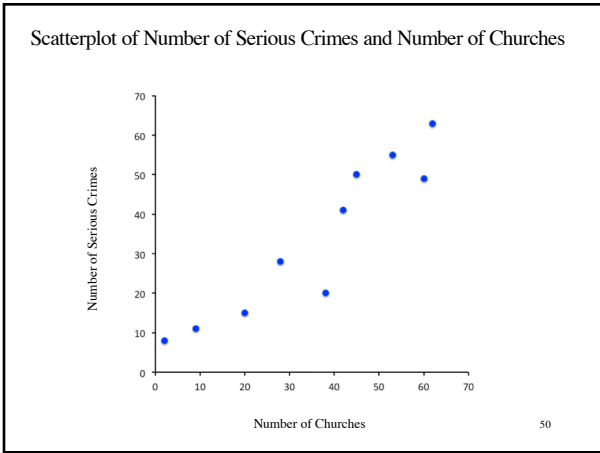
Understanding & Interpreting the Pearson Correlation

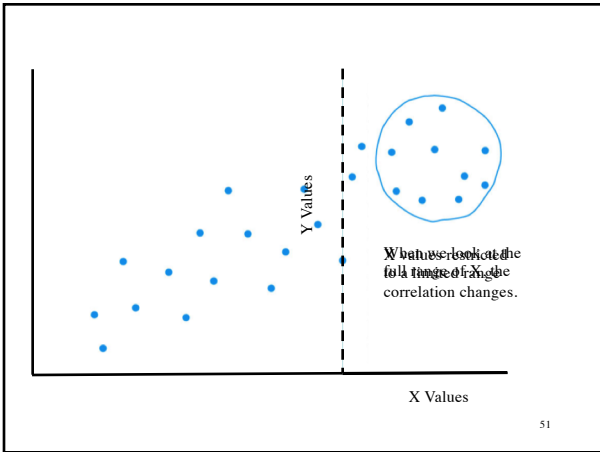
- Correlation is not causation
- Correlation greatly affected by the range of scores represented in the data
- One or two extreme data points (outliers) can dramatically affect the value of the correlation
- How accurately one variable predicts the other—the strength of a relation

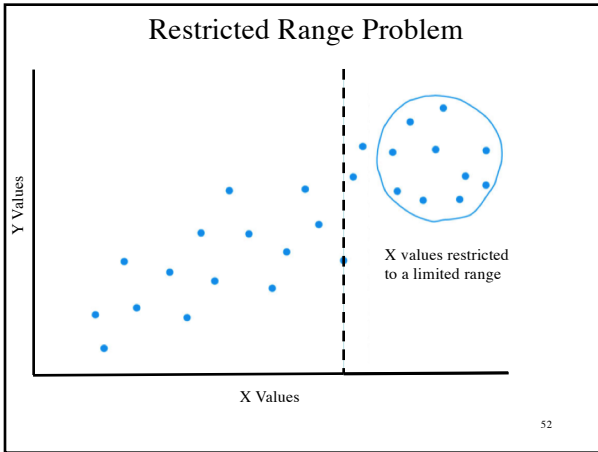
48

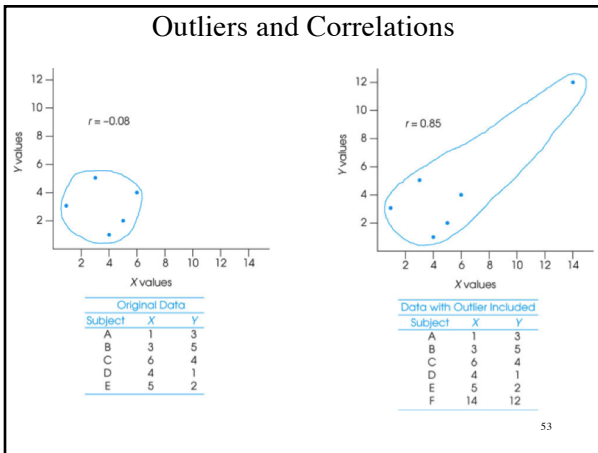
Correlation is not causation

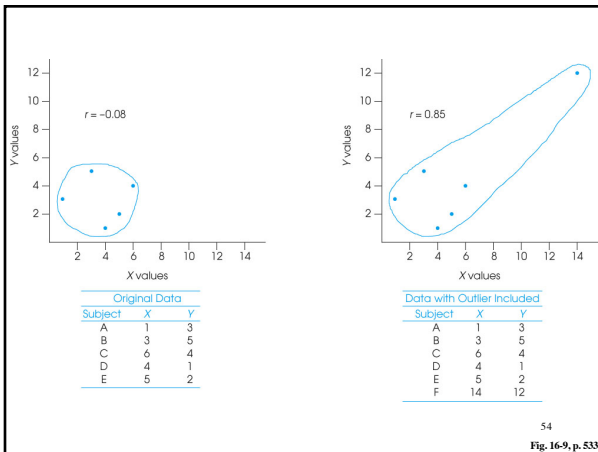
49





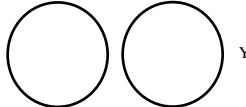




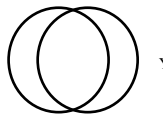


Coefficient of Determination (r^2)

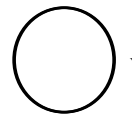
With $r = 0$, None of the Y variability can be predicted from X; $r^2 = 0$



X Y



X Y



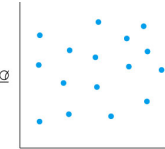
X Y

With $r = 0.8$, the Y variability is partially predicted from the relation with X; $r^2 = 0.64$ or 64%

With $r = 1.0$, the Y variability is completely predicted from the relation with X; $r^2 = 1.00$ or 100%

55

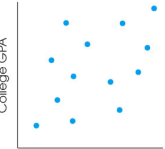
r versus r^2



IQ

Shoe size

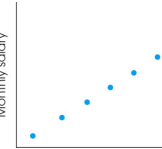
$r = 0$
 $r^2 = 0$



College GPA

IQ

$r = .80$
 $r^2 = .64$ or 64%



Monthly salary

Annual salary

$r = 1.0$
 $r^2 = 1.00$ or 100%

56

Different Types of Correlation

- **Pearson Correlation** – Used only when x and y are **both** interval or ratio scales
- If either X or Y are not interval or ratio scales, then we compute a different correlation coefficient.

57

Other Types of Correlation

- **Spearman Correlation** – Used when either x or y (or both) are ordinal scales
- **Point-Biserial:** Used when one variable is interval or ratio and the other variable is dichotomous (two values – e.g. male/female)
- **Phi-Coefficient** – Used when both variables (x and y) are dichotomous.

58

The Spearman Correlation

- The **Spearman correlation** is used in two general situations:
 - (1) When X and Y are both consist of ranks (ordinal scales).
 - (2) When X and Y are interval/ratio BUT there are outliers in the data--both variables can be converted to ranks and a Spearman correlation is computed.

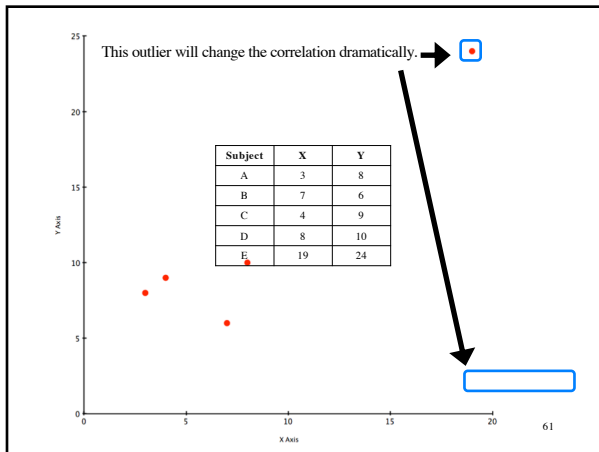
59

The Spearman Correlation (cont.)

The calculation of the Spearman correlation requires:

1. Two variables are observed for each individual.
2. The observations for each variable are rank ordered. The X values and Y values are ranked separately.
3. After the variables have been ranked, the Spearman correlation is computed by using the same formula used for the Pearson but substituting the ranked data.

60



So we transform our scores into ranks...

Original Scores

Subject	X	Y
A	3	8
B	7	6
C	4	9
D	8	10
E	19	24

Convert Original Scores to Ranks

Subject	X Rank	Y Rank
A	3 (1 st)	8 (2 nd)
B	7 (3 rd)	6 (1 st)
C	4 (2 nd)	9 (3 rd)
D	8 (4 th)	10 (4 th)
E	19 (5 th)	24 (5 th)

Compute Correlation on the Ranks

Subject	X Rank	Y Rank
A	1	2
B	3	1
C	2	3
D	4	4
E	5	5

Ranking Tied Scores

- List the scores from smallest to largest. Include tied values in the list.
- Assign a rank (1st, 2nd, 3rd, etc.) to each score in the ordered list.
- When two (or more) scores are tied, compute the mean of their ranks and assign this mean value as the rank for each of the tied scores.

For example...

Here are some scores with ties: 3, 5, 3, 6, 6, 12, 6

1. List the scores from low to high.
2. Assign a rank to each score
3. For each group of tied scores, compute a mean rank and assign it to each tied score in that group.

Scores	Rank	Final Rank
3	1	1.5
3	2	1.5
5	3	3
6	4	5
6	5	5
6	6	5
12	7	7

64

The Point-Biserial Correlation and the Phi Coefficient

- The same correlation formula can also be used to measure the relationship between two variables when one or both of the variables is dichotomous.
- A dichotomous variable is one for which there are exactly two categories: for example, men/women or succeed/fail.

65

The Point-Biserial Correlation and the Phi Coefficient (cont.)

- In situations where one variable is dichotomous and the other consists of regular numerical scores (interval or ratio scale), the resulting correlation is called a **point-biserial correlation**.
- When both variables are dichotomous, the resulting correlation is called a **phi-coefficient**.

66

Point-Biserial Correlation

Used in situations where one variable is measured on an interval or ratio scale but the second variable has only two different values (i.e. a **dichotomous** variable). Examples:

1. Vocabulary scores for Males versus Females
2. IQ scores for college grad versus non college grads
3. Memory scores for visual imagery group versus rote rehearsal group

67

Original Data		Converted Data	
Attitude Score (Y)	Gender (X)	Attitude Score (Y)	Gender (X)
8	Male	8	0
7	Female	7	1
4	Male	4	0
6	Male	6	0
1	Female	1	1
9	Male	9	0
3	Female	3	1
4	Female	4	1

68

The Point-Biserial Correlation (cont.)

- The point-biserial correlation is closely related to the independent-measures t test introduced in Chapter 10.
- When the data consists of one dichotomous variable and one numerical variable, the dichotomous variable can also be used to separate the individuals into two groups.
- Then, it is possible to compute a sample mean for the numerical scores in each group.

69

The Point-Biserial Correlation (cont.)

- In this case, the independent-measures *t* test can be used to evaluate the mean difference between groups.
- The *t*-statistic and the point-biserial correlation coefficient are mathematically related. One can be derived from the other.

70

t^2 and r^2 are related

$$r^2 = \frac{t^2}{t^2 + df} \qquad t^2 = \frac{r^2(df)}{1 - r^2}$$

Remember: $df = n - 2$

71

TABLE 16.3

The same data are organized in two different formats. On the left-hand side, the data appear as two separate samples appropriate for an independent-measures *t* hypothesis test. On the right-hand side, the same data are shown as a single sample, with two scores for each individual: the original high school grade and a dichotomous score (*Y*) that identifies the condition (Sesame Street or not) in which the participant is located. The data on the right are appropriate for a point-biserial correlation.

Data for the Independent-Measures <i>t</i> .				Data for the Point-Biserial Correlation.		
Two separate samples, each with $n = 10$ scores.				Two scores, <i>X</i> and <i>Y</i> , for each of the $n = 20$ participants.		
Average High School Grade				Participant	Grade <i>X</i>	Condition <i>Y</i>
Watched Sesame Street	Did Not Watch Sesame Street					
86	99	90	79	A	86	1
87	97	89	83	B	87	1
91	94	82	86	C	91	1
97	89	83	81	D	97	1
98	92	85	92	E	98	1
				F	99	1
$n = 10$	$n = 10$			G	97	1
$M = 93$	$M = 85$			H	94	1
$SS = 200$	$SS = 160$			I	89	1
				J	92	1
				K	90	0
				L	89	0
				M	82	0
				N	83	0
				O	85	0
				P	79	0
				Q	83	0
				R	86	0
				S	81	0
				T	92	0

72

Phi Coefficient

Original Data		Converted Data	
Birth Order (X)	Personality (Y)	Birth Order (X)	Personality (Y)
1 st	Introvert	0	0
3 rd	Extrovert	1	1
Only	Extrovert	0	1
y	Extrovert	1	1
2 nd	Extrovert	1	1
4 th	Introvert	1	0
2 nd	Introvert	0	0
Only	Extrovert	1	1
y			
3 rd			
