Graduate Statistics

Correlation and Regression

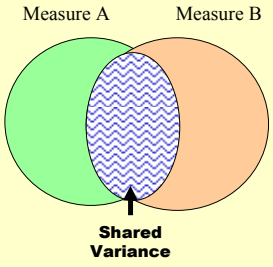
What We Will Cover in This Section

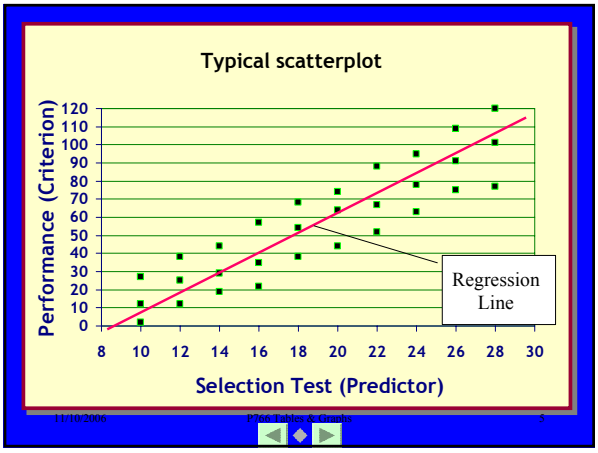
- Overview
- Correlation
- Regression

What Correlational Statistics Do

1. Assess the strength of the relationship between two or more variables.
2. Determine the direction of the relationship.
   - Positive.
   - Negative.

Correlational Model

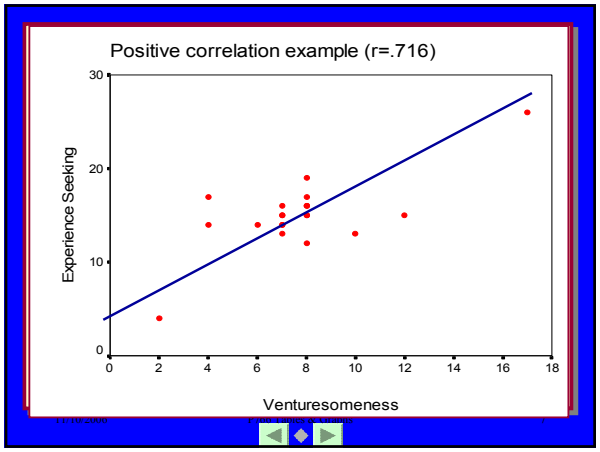




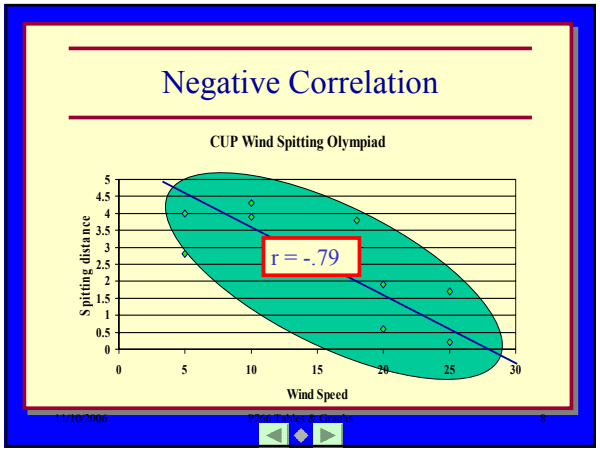
The Null Hypothesis

There is no relationship between the variables

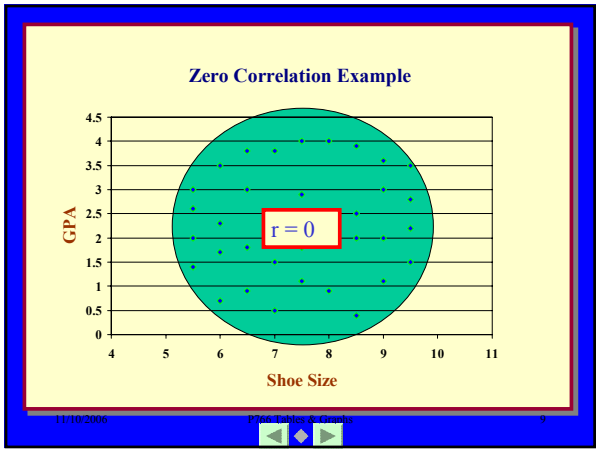
$\rho = 0$



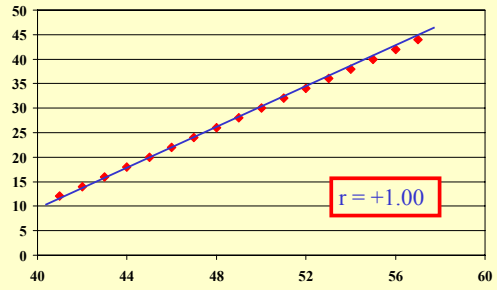
Negative Correlation



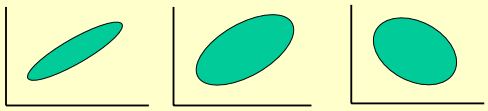
Zero Correlation Example



Perfect Positive Correlation Example



Relationship Strength



Strong **Moderate** **Weak**

The Correlation Coefficient

- Indicated by r.
- Ranges from –1.00 to +1.00
  - The number indicates the strength of the relationship.
  - The sign indicates whether the relationship is positive or negative.
- Does NOT indicate causality.

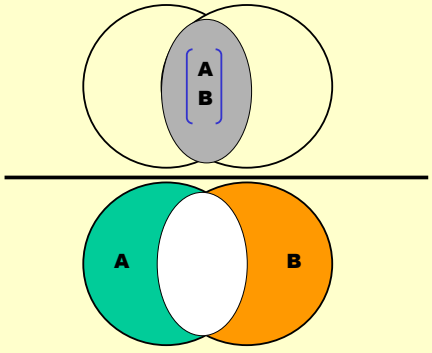
Where Simple Correlations are Used

1. Prediction.
2. Validation studies.
3. Reliability studies.
4. Theoretical studies.
5. Identification of surrogate variables.

General Model

$$r = \frac{\text{Degree to which X and Y covary}}{\text{Variability of X and Y separately}}$$

$$r = \frac{\text{Shared variability}}{\text{Variability of X and Y separately}}$$



Definitional Formula

$$r = \frac{\Sigma\left(Z_x \text{ x } Z_y\right)}{N}$$

$$r = \frac{N(\sum_{i=1}^{n} X_i Y_i) - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y)_i}{\sqrt{N\left(\sum_{i=1}^{n} X_i^2\right) - \left(\sum_{i=1}^{n} X_i\right)^2 \text{ x } N\left(\sum_{i=1}^{n} Y_i^2\right) - \left(\sum_{i=1}^{n} Y_i\right)^2}}$$

An equation you don't need to know.

The Significance Test for r

- H_o: $\rho = 0$
- H_a: $\rho \neq 0$
- df = N – 2

  Note: N is the number of **pairs** of scores.
- See Table B6, page 709.

	Height	Hand	Zheight	Zhand	
	70	9.25	1.59	1.81	2.88
	65	8	-0.02	0.32	-0.01
	66	7.75	0.30	0.02	0.01
	70	8.25	1.59	0.62	0.98
	63	8	-0.66	0.32	-0.21
	62	6	-0.98	-2.08	2.04
	62.5	7.75	-0.82	0.02	-0.01
	61	7	-1.30	-0.88	1.15
	66	7.625	0.30	-0.13	-0.04
Mean	65.06	7.74			0.75
SD	3.11	0.83			

r and Effect Size

Coefficient of Determination: r^2

Represents the amount of **shared variance** between the two variables.

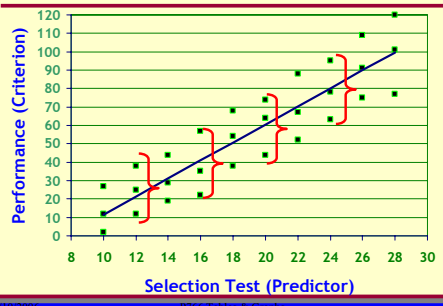
Coefficient of Alienation: $1 - r^2$

Represents the amount of variance that is **not shared**.

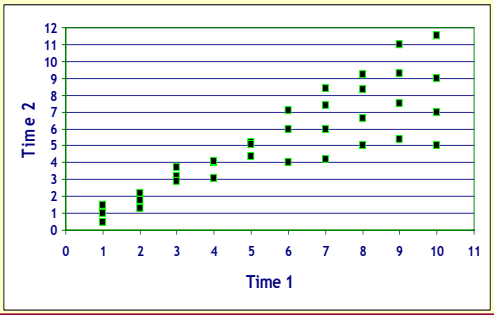
Assumptions When Using the Pearson Correlation Coefficient

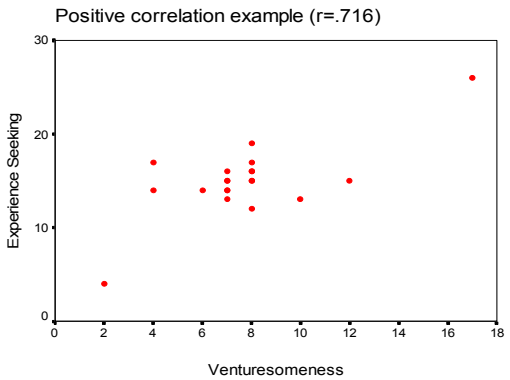
1. The relationship is linear.
2. Interval or ratio measurement scale.
3. Homoscedasticity.
4. The distribution is continuous.
5. There is no restriction of range.
6. The distribution is normal.

Assumption of Homoscedasticity



Violation of Homoscedasticity



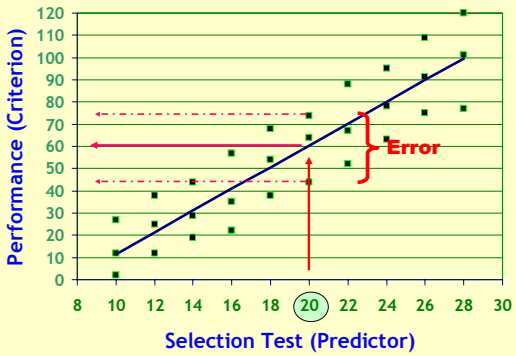


Consequences of Restriction of Range

	Unrestricted		Restricted	
	Experience Seeking	Venture-someness	Experience Seeking	Venture-someness
Mean	15.00	7.65	15.00	7.44
Standard Deviation	3.90	3.05	1.68	**1.82**
Correlation	r = .716		r = -.058	

Regression

Statistical technique using the correlation coefficient to make predictions about one variable based on its relationship with one or more other variables.



Regression Formula

$$\hat{Z}_y = r(Z_x)$$

Residual

This is the error in prediction and represents the amount of variance not accounted for in prediction.

Standard Error of Estimate (SE_{est})

$$SE_{est} = \sqrt{\frac{\Sigma\left(\hat{Y} - Y\right)^2}{N - 1}}$$

Represents the standard deviation of the error in predicting Y from X in regression.

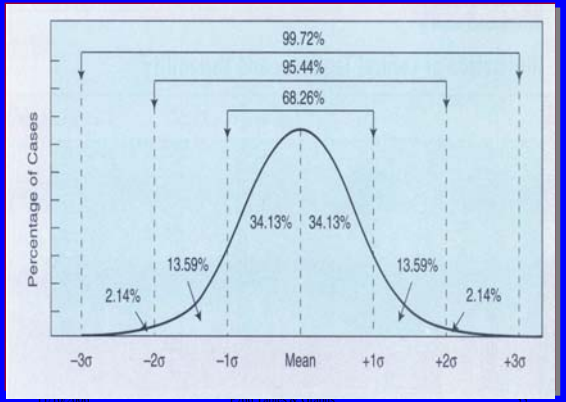
SE_{est} Computational Formula

$$SE_{est(y)} = SD_y\sqrt{1-r_{xy}^{\ 2}}$$

$$SE_{est(x)} = SD_x\sqrt{1-r_{xy}^{\ 2}}$$

Use of the Standard Error of Estimate

- When r^2 does not equal 1.00 there will be some error in our predictions.
- The standard deviation in the error of our predictions is represented by SE_{est}.
- SE_{est} gives a way to determine the range where a predicted score will most likely fall and the probability that the score will fall within this range.

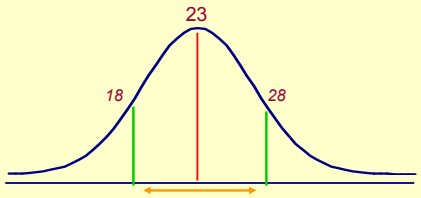


Confidence Intervals

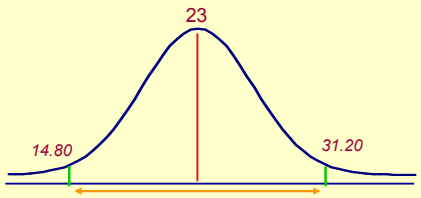
Computing the confidence intervals when Y' = 23 , r=.70, and SD_y= 7

Confidence Interval	SE_{est}	±(SE_{est} x Z)	Range
68% ±1.00Z	± 5.00	± 5.00	18.00 - 28.00
90% ±1.64Z	± 5.00	± 8.20	14.80 - 31.20
99% ±2.58Z	± 5.00	± 12.90	10.10 – 35.90

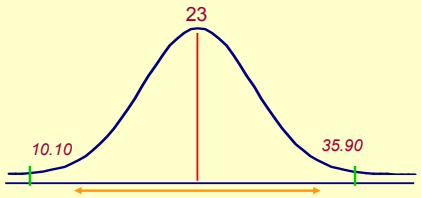
68% Confidence Interval (±1.00Z)



90% Confidence Interval (± 1.64 Z)



99% Confidence Interval (±2.58Z)



Influence of Increasing r on SE_{est}

(SD = 10)

r	r^2	SE_{est}
.90	.81	4.35
.80	.64	6.0
.70	.49	7.1
.60	.36	8.0

Key Learning Points

1. The correlation coefficient assesses the strength and direction of the relationship between two or more variables.
2. The correlation coefficient is sensitive to the variability in the sample.
3. One cannot infer causality from examining the correlation coefficient.
4. The effect size can be assessed by r^2.

Key Learning Points, cont.

5. There will always be error in predicting one variable from another.
6. The degree of error is directly related to the strength of the correlation between the two variables.