Chapter 10

Correlation and Regression

Chapter 10
Correlation and Regression

Overview
Paired Data

- is there a relationship
- if so, what is the equation
- use the equation for prediction

Example: Lengths and Weights of Male Bears

x Length (in.)	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
y Weight (lb)	80	344	416	348	262	360	332	34

10-2

Correlation

Definition

- **Correlation**

  exists between two variables when one of them is related to the other in some way

Definition

❖ **Scatterplot (or scatter diagram) is a graph in which the paired (*x*,*y*) sample data are plotted with a horizontal *x* axis and a vertical *y* axis. Each individual (*x*,*y*) pair is plotted as a single point.**

Example: Lengths and Weights of Male Bears

x Length (in.)	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
y Weight (lb)	80	344	416	348	262	360	332	34

Example: Lengths and Weights of Male Bears

x Length (in.)	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
	↓	↓	↓	↓	↓	↓	↓	↓
y Weight (lb)	80	344	416	348	262	360	332	34

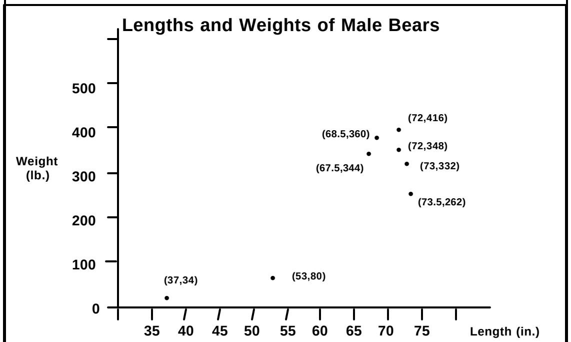
Example: Lengths and Weights of Male Bears

x Length (in.)	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
	↓	↓	↓	↓	↓	↓	↓	↓
y Weight (lb)	80	344	416	348	262	360	332	34

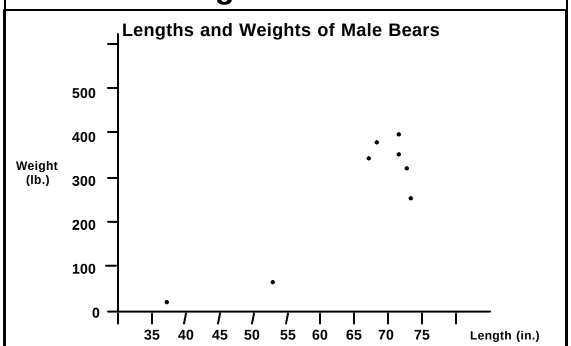
(x , y) = (Length , Weight)

(53.0, 80)
(67.5, 344)
(72.0, 416)
etc.

Scatter Diagram of Paired Data



Scatter Diagram of Paired Data



Positive Linear Correlation

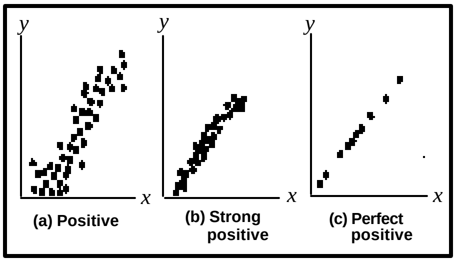


Figure 10-2 **Scatter Plots**

Negative Linear Correlation

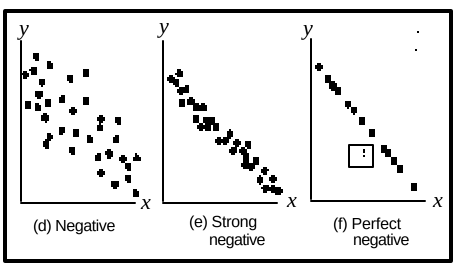


Figure 10-2 **Scatter Plots**

No Linear Correlation

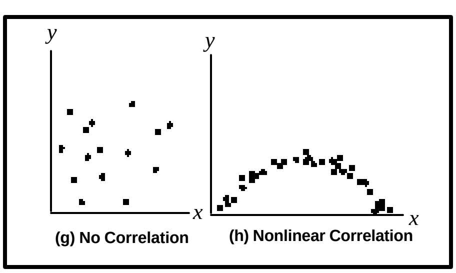


Figure 10-2 **Scatter Plots**

Definition

❖ **Linear Correlation Coefficient *r***

measures <u>strength</u> of the linear relationship between paired x-and y-quantitative values in a sample

Definition

❖ **Linear Correlation Coefficient *r***

sometimes referred to as the Pearson product moment correlation coefficient

Assumptions

1. **The sample of paired data (x,y) is a <u>random</u> sample.**
2. **The pairs of (x,y) data have a bivariate normal distribution.**

Notation for the Linear Correlation Coefficient

- *n* — number of pairs of data presented.
- S — denotes the addition of the items indicated.
- Sx — denotes the sum of all x values.
- Sx^2 — indicates that each x score should be squared and then those squares added.
- (Sx)^2 — indicates that the x scores should be added and the total then squared.
- Sxy — indicates that each x score should be first multiplied by its corresponding y score. After obtaining all such products, find their sum.
- ***r*** — represents linear correlation coefficient for a <u>sample</u>
- *r* — represents linear correlation coefficient for a <u>population</u>

Definition
Linear Correlation Coefficient *r*

$$r = \frac{nSxy - (Sx)(Sy)}{\sqrt{n(Sx^2) - (Sx)^2}\,\sqrt{n(Sy^2) - (Sy)^2}}$$

Formula 10-1

Definition
Linear Correlation Coefficient *r*

$$r = \frac{nSxy - (Sx)(Sy)}{\sqrt{n(Sx^2) - (Sx)^2}\,\sqrt{n(Sy^2) - (Sy)^2}}$$

Formula 10-1

<u>Calculators can compute *r*</u>

r (rho) is the linear correlation coefficient for <u>all</u> paired data in the population.

Rounding the Linear Correlation Coefficient *r*

- Round to three decimal places so that it can be compared to critical values in Table A-5
- Use calculator or computer if possible

Example: Lengths and Weights of Male Bears

x Length (in.)	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
y Weight (lb)	80	344	416	348	262	360	332	34

Example: Lengths and Weights of Male Bears

x Length (in.)	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
y Weight (lb)	80	344	416	348	262	360	332	34

r = 0.897 using your calculator

Interpreting the Linear Correlation Coefficient

- ❖ **If the absolute value of *r* exceeds the value in Table A - 5, conclude that there is a significant linear correlation.**
- ❖ **Otherwise, there is not sufficient evidence to support the conclusion of significant linear correlation.**

TABLE A-5 Critical Values of the Pearson Correlation Coefficient r

n	*a* = .05	*a* = .01
4	.950	.999
5	.878	.959
6	.811	.917
7	.754	.875
8	.707	.834
9	.666	.798
10	.632	.765
11	.602	.735
12	.576	.708
13	.553	.684
14	.532	.661
15	.514	.641
16	.497	.623
17	.482	.606
18	.468	.590
19	.456	.575
20	.444	.561
25	.396	.505
30	.361	.463
35	.335	.430
40	.312	.402
45	.294	.378
50	.279	.361
60	.254	.330
70	.236	.305
80	.220	.286
90	.207	.269
100	.196	.256

Properties of the Linear Correlation Coefficient *r*

1. **-1 £ *r* £ 1**
2. **Value of *r* does not change if all values of either variable are converted to a different scale.**
3. **The value of *r* is not affected by the choice of *x* and *y*. Interchange *x* and *y* and the value of *r* will not change.**
4. ***r* measures strength of a <u>linear</u> relationship.**

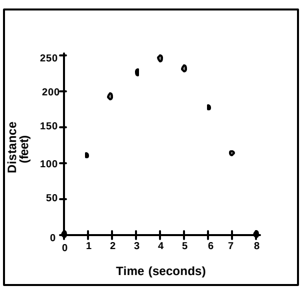
Common Errors Involving Correlation

1. **Causation: It is incorrect to conclude that correlation implies causality.**

2. **Averages: Averages suppress individual variation and may inflate the correlation coefficient.**

3. **Linearity: There may be <u>some relationship</u> between x and y even when there is no significant linear correlation.**

Common Errors Involving Correlation



Scatterplot of Distance above Ground and Time for Object Thrown Upward

Formal Hypothesis Test

- **To determine whether there is a significant linear correlation between two variables**
- **Two methods**
- **Both methods let** $H_0: r = 0$
  (no significant linear correlation)
  $H_1: r \neq 0$
  (significant linear correlation)

Method 1: Test Statistic is *t*

(follows format of earlier chapters)

Test statistic:

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Method 1: Test Statistic is *t*

(follows format of earlier chapters)

Test statistic:

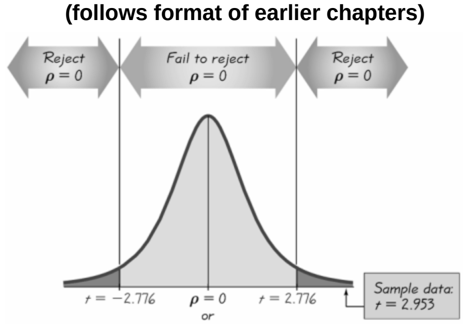
$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Critical values:

use Table A-3 with degrees of freedom = $n - 2$

Method 1: Test Statistic is *t*

(follows format of earlier chapters)



Method 2: Test Statistic is *r*

(uses fewer calculations)

- **Test statistic:** ***r***
- **Critical values: Refer to Table A-5**

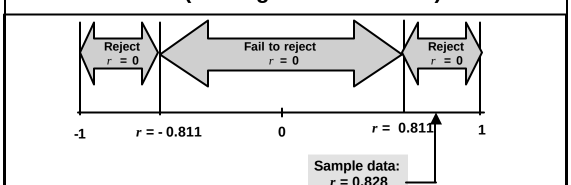
  (no degrees of freedom)

Method 2: Test Statistic is *r*

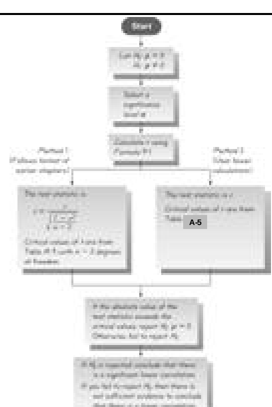
(uses fewer calculations)

- **Test statistic:** ***r***
- **Critical values: Refer to Table A-5**

  (no degrees of freedom)



Testing for a Linear Correlation



Is there a significant linear correlation?

Data from the Garbage Project

x Plastic (lb)	0.27	1.41	2.19	2.83	2.19	1.81	0.85	3.05
y Household	2	3	3	6	4	2	1	5

Is there a significant linear correlation?

Data from the Garbage Project

x Plastic (lb)	0.27	1.41	2.19	2.83	2.19	1.81	0.85	3.05
y Household	2	3	3	6	4	2	1	5

$n = 8$ $\alpha = 0.05$ $H_0: \rho = 0$

$H_1: \rho \neq 0$

Test statistic is $r = 0.842$

Is there a significant linear correlation?

$n = 8$ $\alpha = 0.05$ $H_0: \rho = 0$

$H_1: \rho \neq 0$

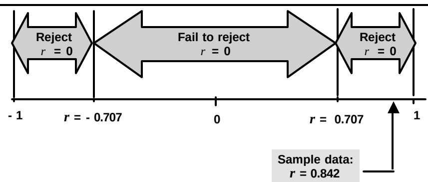
Test statistic is $r = 0.842$

Critical values are $r = -0.707$ and 0.707 (Table A-5 with $n = 8$ and $\alpha = 0.05$)

n	$\alpha = .05$	$\alpha = .01$
4	.950	.999
5	.878	.959
6	.811	.917
7	.754	.875
8	.707	.834
9	.666	.798
10	.632	.765
11	.602	.735
12	.576	.708
13	.553	.684
14	.532	.661
15	.514	.641
16	.497	.623
17	.482	.606
18	.468	.590
19	.456	.575
20	.444	.561
25	.396	.505
30	.361	.463
35	.335	.430
40	.312	.402
45	.294	.378
50	.279	.361
60	.254	.330
70	.236	.305
80	.220	.286
90	.207	.269
100	.196	.256

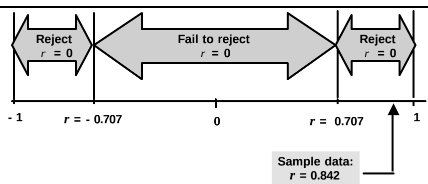
TABLE A-5 Critical Values of the Pearson Correlation Coefficient r

Is there a significant linear correlation?



Is there a significant linear correlation?

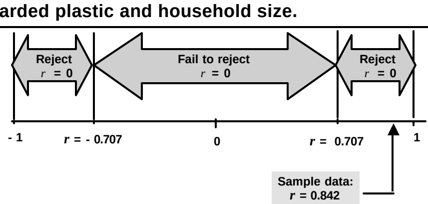
0.842 > 0.707
The test statistic <u>does</u> fall within the critical region.



Is there a significant linear correlation?

0.842 > 0.707
The test statistic <u>does</u> fall within the critical region.

Therefore, we REJECT H_0: $r = 0$ (no correlation) and conclude <u>there is</u> a significant linear correlation between the weights of discarded plastic and household size.



Justification for *r* Formula

Justification for *r* Formula

Formula 10-1 is developed from

$$r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{(n-1) S_x S_y}$$

Justification for *r* Formula

Formula 10-1 is developed from

$$r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{(n-1) S_x S_y}$$

$(\bar{x}, \bar{y})$ centroid of sample points

Justification for *r* Formula

Formula 10-1 is developed from

$$r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{(n-1) S_x S_y}$$

$(\bar{x}, \bar{y})$ **centroid of sample points**