Curtin College

DIPLOMA OF INFORMATION TECHNOLOGY

RNI1006 REGRESSION AND NONPARAMETRIC INFERENCE

COMMONWEALTH OF AUSTRALIA

Copyright Regulations 1969

WARNING

This material has been reproduced and communicated to you or on behalf of **Curtin College** *pursuant to Part VB of the Copyright Act 1968 (the Act).*

The material in this communication may be subject to copyright under the Act. Any further reproduction or communication of this material by you may be the subject of copyright protection under the ACT.

Do not remove this notice.

Acknowledgement

We respectfully acknowledge the Elders and custodians of the Whadjuk Nyungar nation, past and present, their descendants and kin. Curtin College Bentley Campus enjoys the privilege of being located in Whadjuk / Nyungar Boodjar (country) on the site where the Derbal Yerrigan (Swan River) and the Djarlgarra (Canning River) meet. The area is of great cultural significance and sustains the life and well being of the traditional custodians past and present.

Aims of Lecture Week 5

Aim 1 Examining relationships (Moore et 2021, Ch 2; ActEd CS1 Ch 11)

- Response Variables and Explanatory Variables

Aim 2 Relationship Between a Continuous Response Variable (y) and a Continuous Explanatory Variable (x) (Moore et 2021, Ch 2; ActEd CS1 Ch 11)

2.1 Graphically – Scatterplot

2.2 Numerically – Correlation (Pearson, parametric) (BREAK 5 mins)

2.3 Hypothesis Testing - Correlation

Aim 3 Other types of correlation coefficients (Hollander and Wolfe 2014; ActEd CS1 Ch 12)

3.1 Spearman (nonparametric)

3.2 Kendall (nonparametric)

The BIG picture:
Aim 1 Examining Relationships

Most statistical studies involve more than one variable.

Questions:

- What cases does the data describe?
- What variables are present and how are they measured?
- Are all of the variables quantitative?
- Do some of the variables explain or even cause changes in other variables?

Aim 1 Relationship between two numerical variables

- **Example 1**. Here, we have two quantitative variables for each of 16 students.
  1) How many beers they drank, and
  2) Their blood alcohol level (BAC)

- We are interested in the relationship between the two variables: How is one affected by changes in the other one?

Student	Beers	Blood Alcohol
1	5	0.1
2	2	0.03
3	9	0.19
6	7	0.095
7	3	0.07
9	3	0.02
11	4	0.07
13	5	0.085
4	8	0.12
5	3	0.04
8	5	0.06
10	5	0.05
12	6	0.1
14	7	0.09
15	1	0.01
16	4	0.05

Looking at relationships

- Start with a graph

- Look for an overall pattern and deviations from the pattern

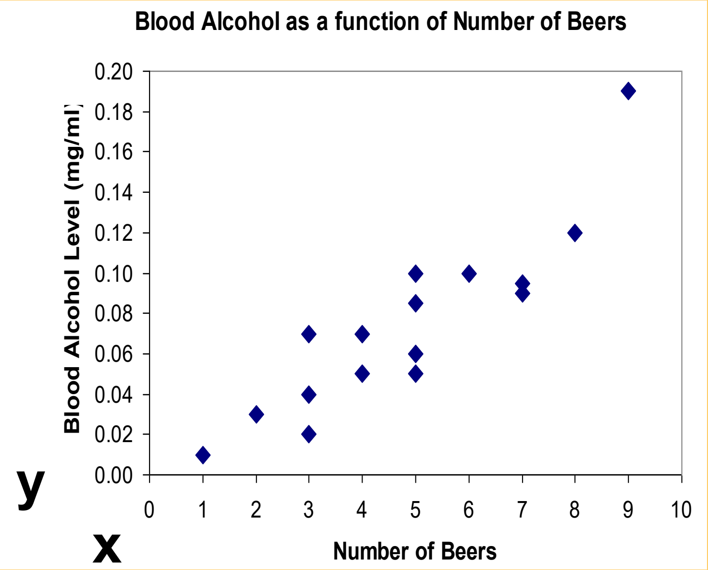
- Use numerical descriptions of the data and overall pattern (if appropriate)

Explanatory and response variables

- A **response variable** measures or records an outcome of a study.
- An **explanatory variable** explains changes in the response variable.
- Typically, the *explanatory* or *independent variable* is plotted on the *x* axis, and the *response* or *dependent variable* is plotted on the *y* axis.

Example 1

Response (dependent) variable: *blood alcohol content*



Explanatory (independent) variable: *number of beers*

Relationships involving numerical variables

- (Y, response) Numerical variable and (X, explanatory) Categorical variable

  Side by side box-plots

  Alternatively, histograms

- (Y, response) Numerical variable and (X, explanatory) Numerical variable

  Scatterplot

Exploring data to find possible relationships

- Different plots for different combinations of types of variables

- Example 2 Two example plots on the golf ball data. The data columns are
  - Brand
  - Distance (of flight of golf ball when hit by a robotic club)
  - Durability measure

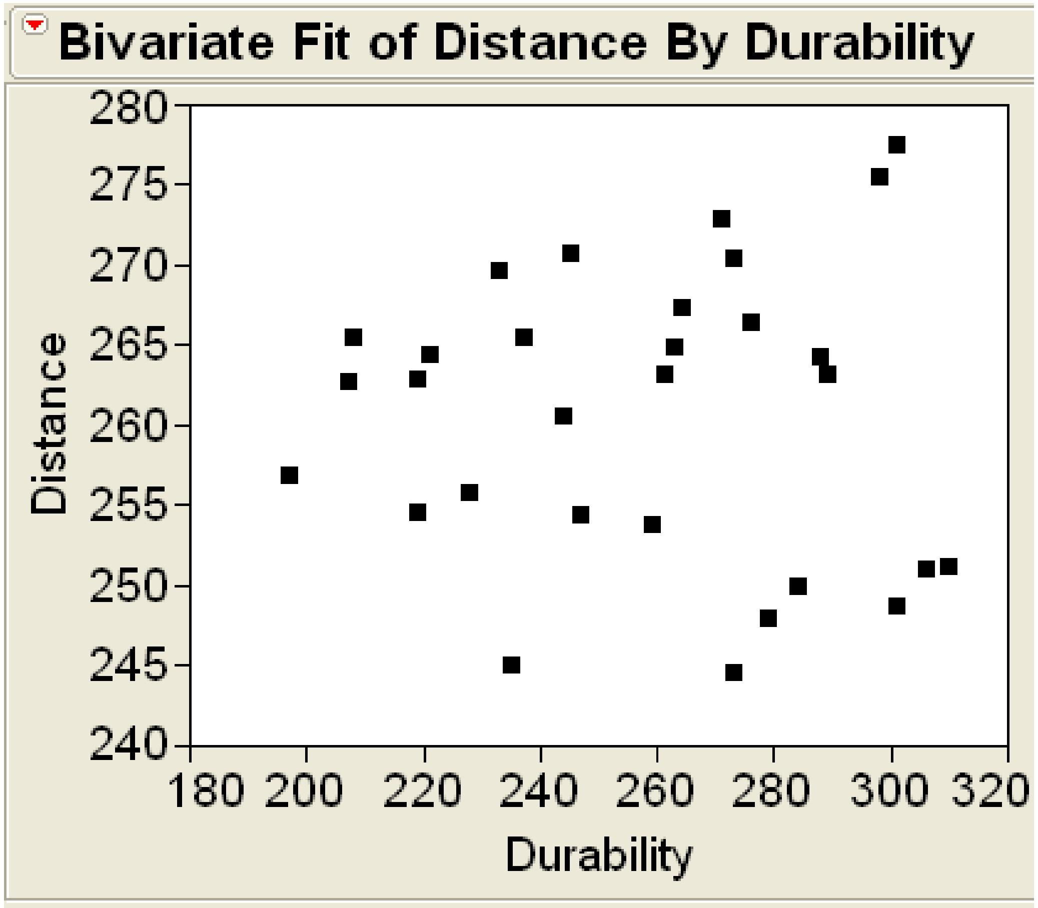
Example 2
Golf ball data

	Brand	Distance	Durability
1	Brand A	251.2	310
2	Brand B	263.2	261
3	Brand C	269.7	233
4	Brand A	245.1	235
5	Brand B	262.9	219
6	Brand C	263.2	289
7	Brand A	248.0	279
8	Brand B	265.0	263
9	Brand C	277.5	301
10	Brand A	251.1	306
11	Brand B	254.5	247
12	Brand C	267.4	264
13	Brand A	265.5	237
14	Brand B	264.3	288
15	Brand C	270.5	273
16	Brand A	250.0	284
17	Brand B	257.0	197
18	Brand C	265.5	208
19	Brand A	253.9	259
20	Brand B	262.8	207
21	Brand C	270.7	245
22	Brand A	244.6	273
23	Brand B	264.4	221
24	Brand C	272.9	271
25	Brand A	254.6	219
26	Brand B	260.6	244
27	Brand C	275.6	298
28	Brand A	248.8	301
29	Brand B	255.9	228
30	Brand C	266.5	276

In-Class Exercise 1.
What are the data types for Brand, Distance and Durability?

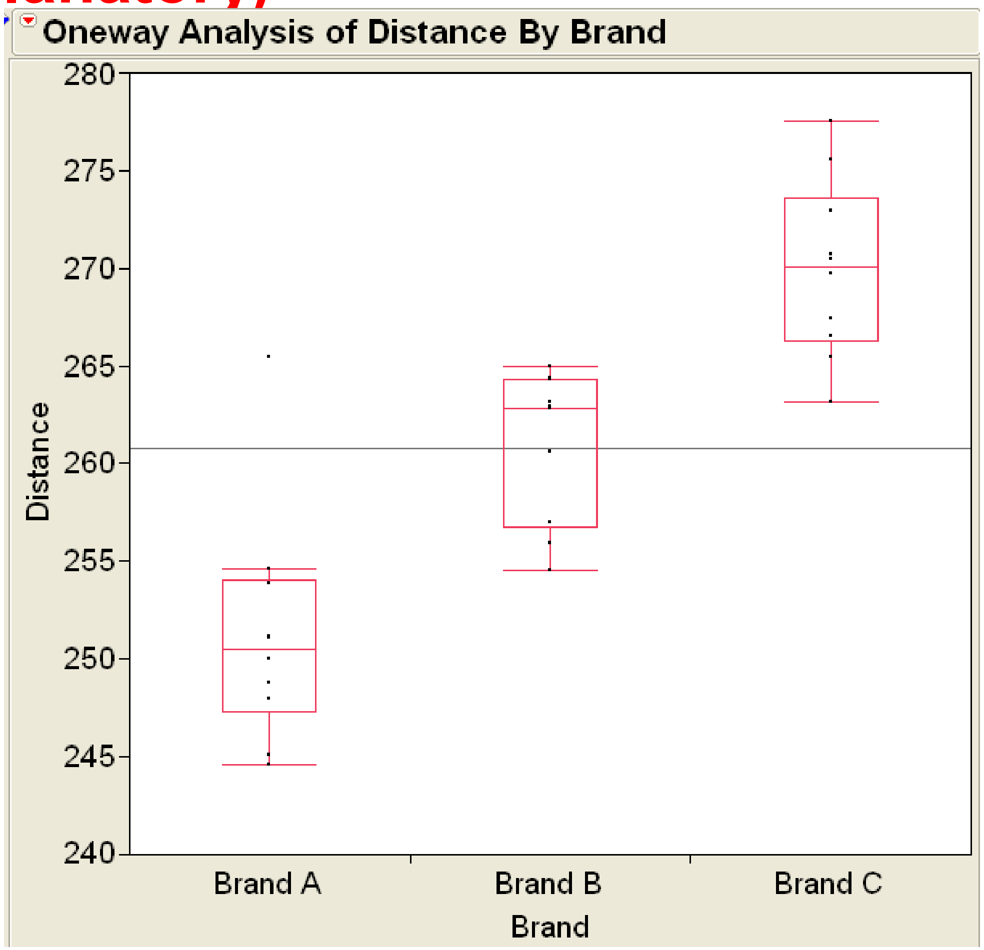
Example 2: Golf Ball (continued)

Distance **(numerical response) explained by *Durability*** **(numerical explanatory)**



R: plot(x,y,….)

Example 2 (ctnd) Distance (numerical response) explained by *Brand* (categorical explanatory)



In Class Exercise 2.
Start with 3S by comparing:

- Shape?
- (S) Center
- Spread?

R: boxplot(y ~ x,........)

Now, if you have two numerical variables……

Aim 2 Linear relationship between two numerical variables

Relationship between two numerical variables can be summarised:

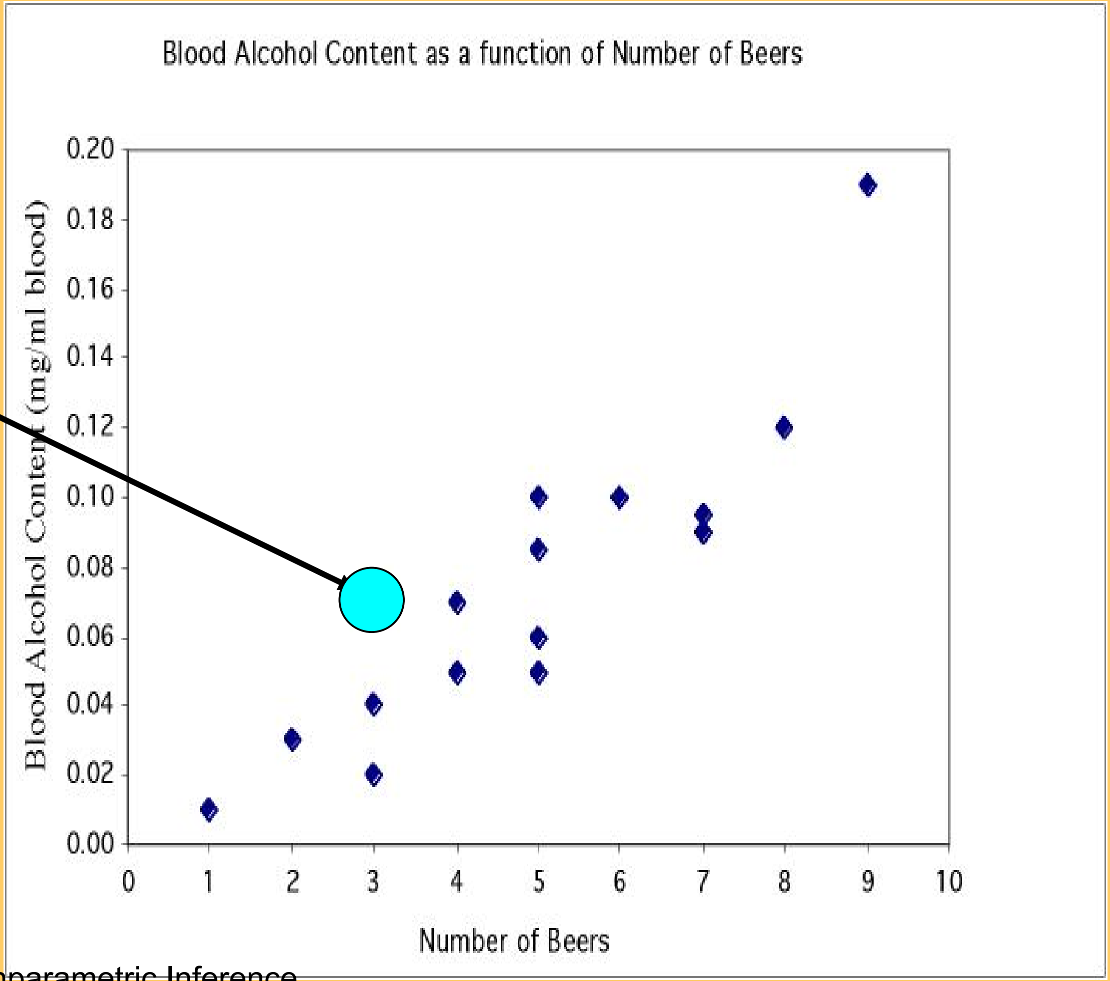
Graphically ⇨ - Scatterplot

Numerically ⇨ - Correlation

Aim 2.1 Scatterplots

In a **scatterplot**, one axis is used to represent each of the variables, and the data are plotted as points on the graph.

Student	Beers	BAC
1	5	0.1
2	2	0.03
3	9	0.19
6	7	0.095
7	3	0.07
9	3	0.02
11	4	0.07
13	5	0.085
4	8	0.12
5	3	0.04
8	5	0.06
10	5	0.05
12	6	0.1
14	7	0.09
15	1	0.01
16	4	0.05



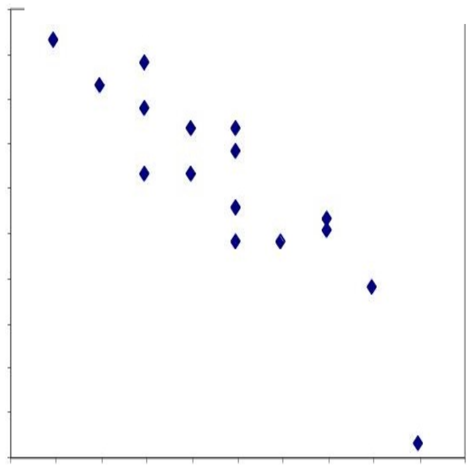
Interpreting scatterplots

- After plotting two variables on a scatterplot, we describe the relationship by examining the **form, direction,** and **strength** of the association.
- We look for an **overall pattern** …
    - **Form:** linear, curved, clusters, no pattern
    - **Direction:** positive, negative, no direction
    - **Strength:** how closely the points fit the “form”
- … and deviations from that pattern: **Outliers**

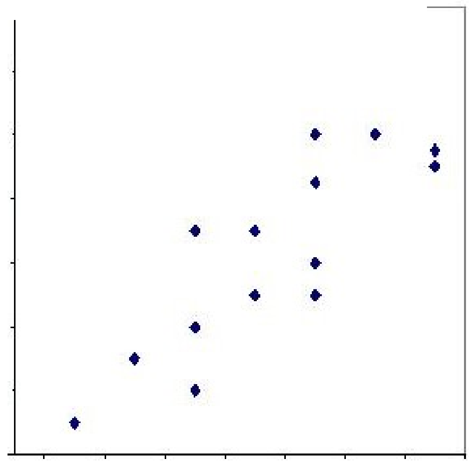
Form and direction of an association

Linear

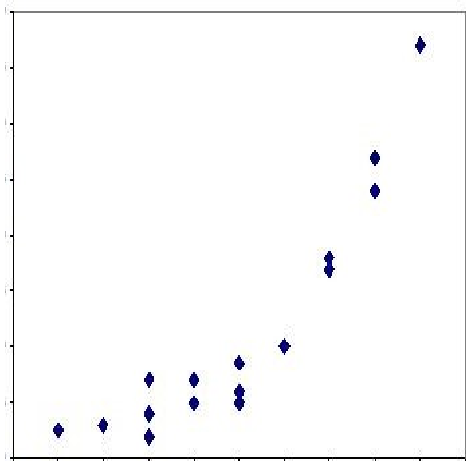
Negative

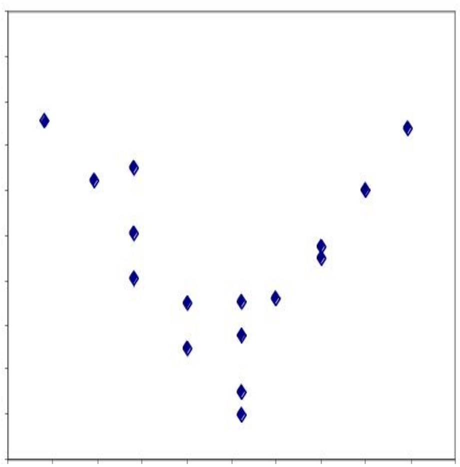


Positive

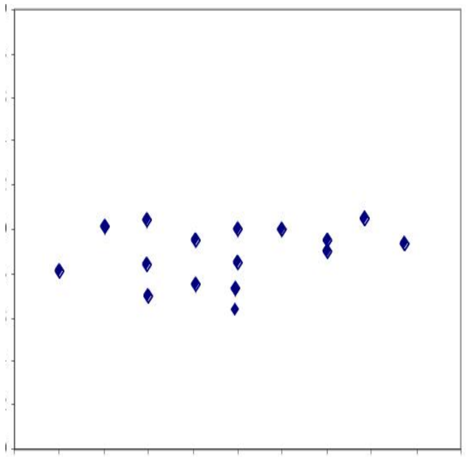


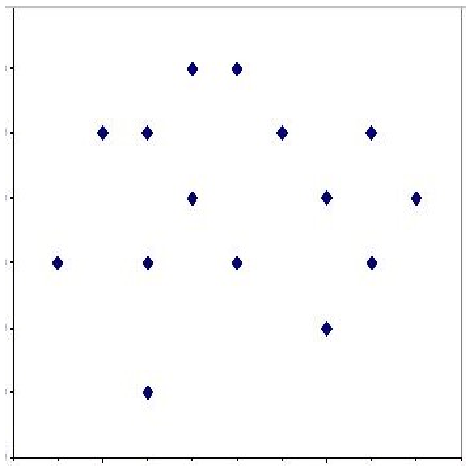
Nonlinear





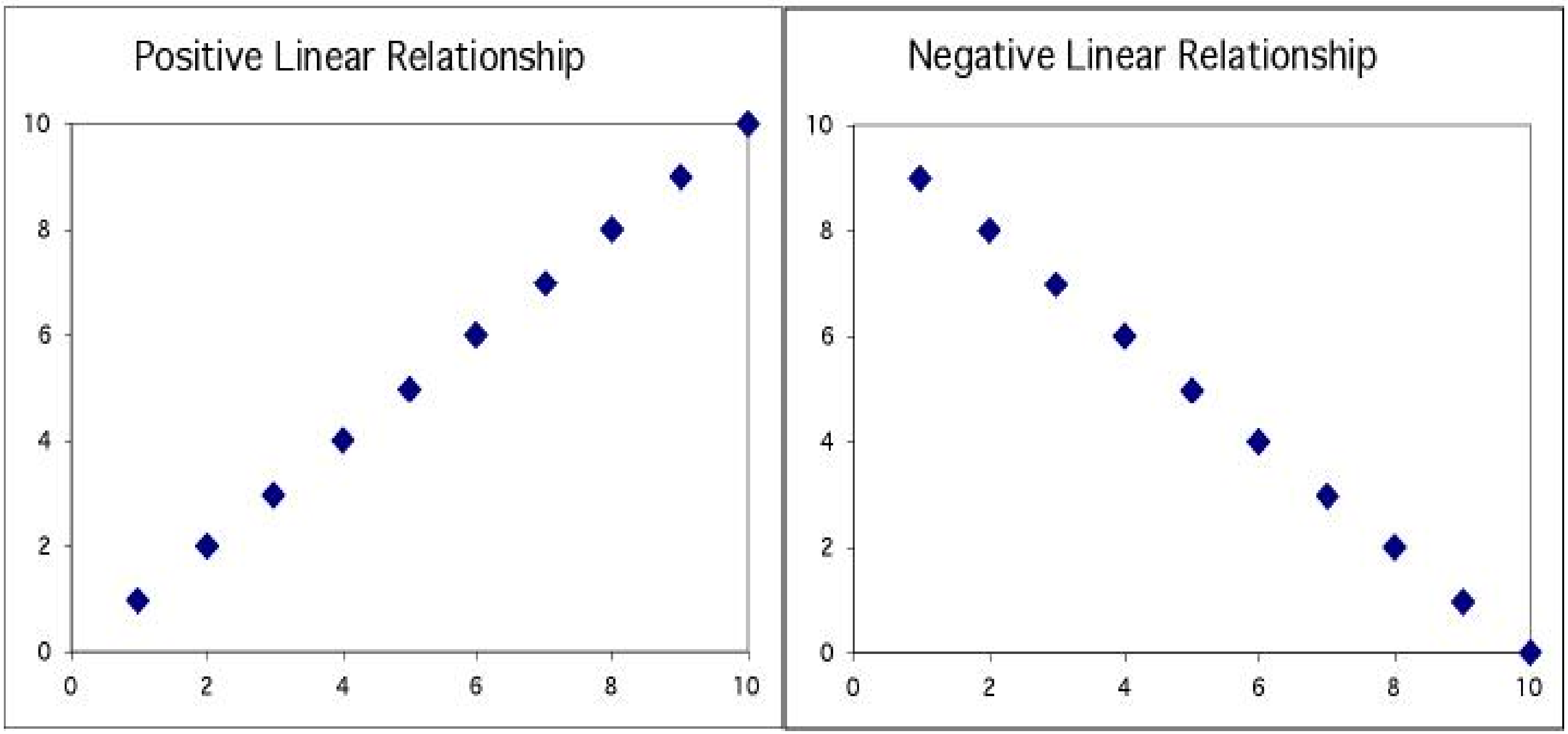
No relationship



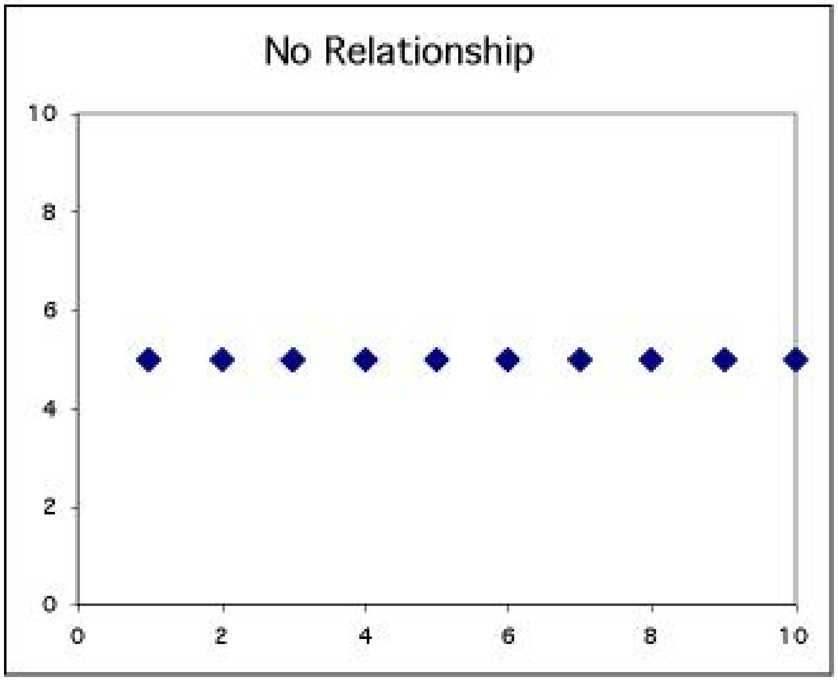


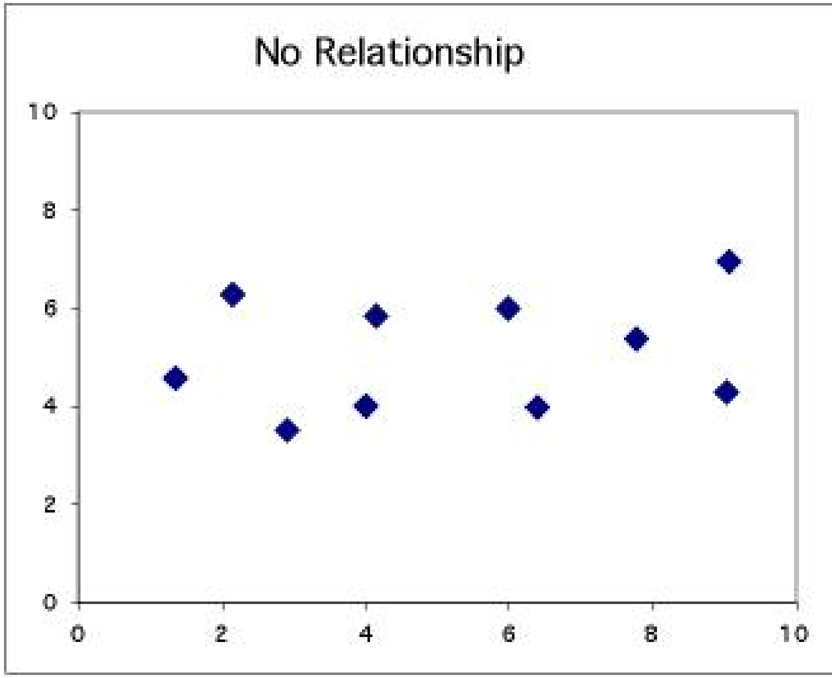
Positive direction/association: High values of one variable tend to occur together with high values of the other variable.

Negative direction/association: High values of one variable tend to occur together with low values of the other variable.



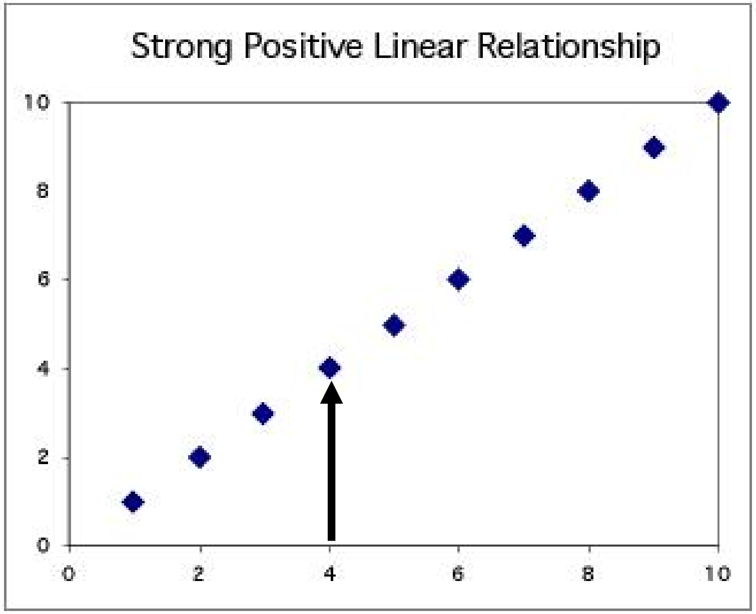
No relationship: X and Y vary independently. Knowing X tells you nothing about Y.

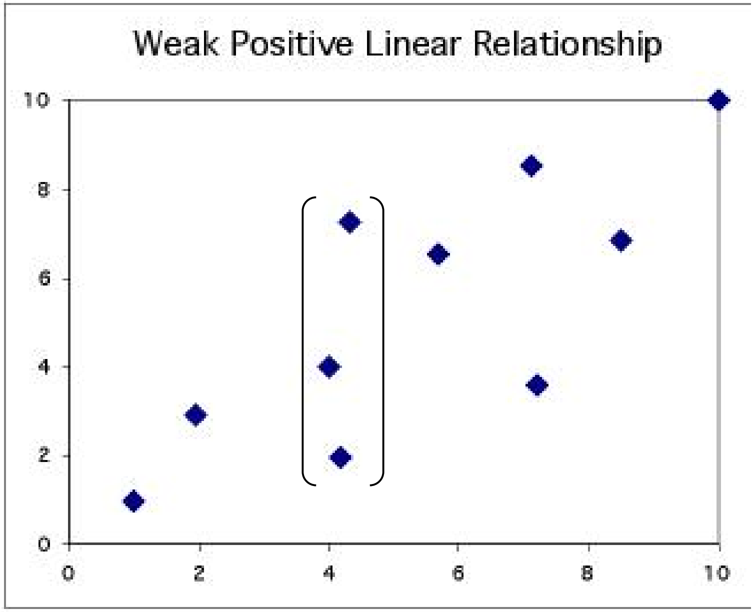




Strength of the association

The **strength** of the relationship between the two variables can be seen by **how much variation**, or **scatter,** there is around the **main form.**

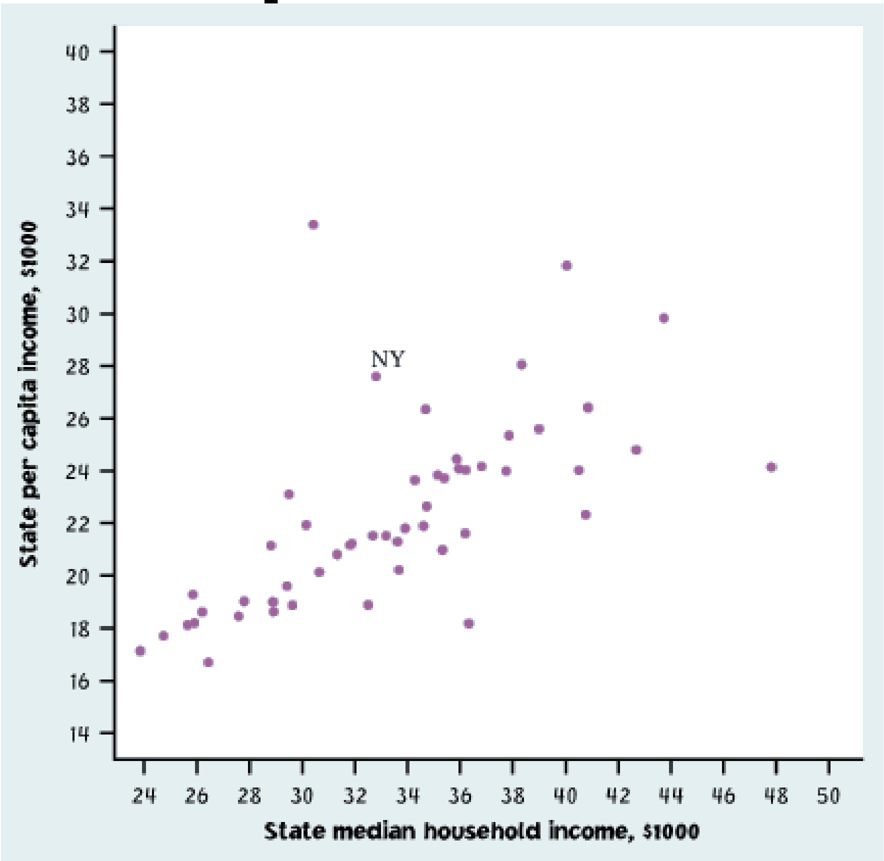


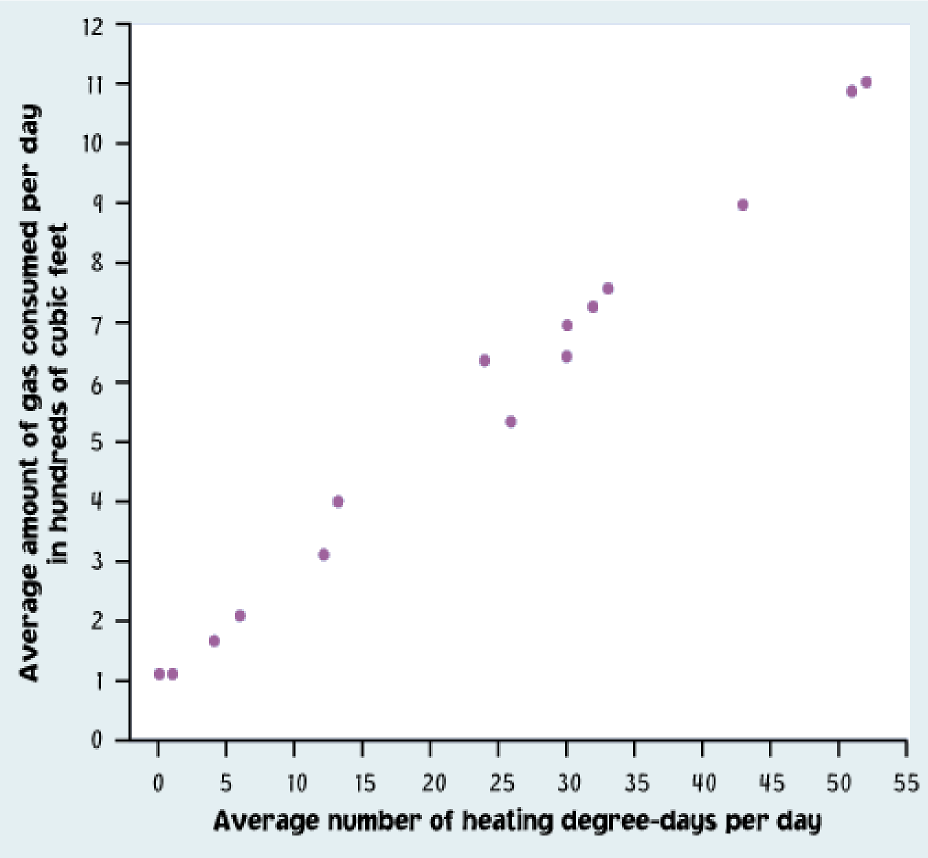


With a strong relationship, you can get a pretty good estimate of *y* if you know *x.*

With a weak relationship, for any *x* you might get a wide range of *y* values.

Example 3





This is a **weak** relationship. For a particular state median household income, you can't predict the state per capita income very well.

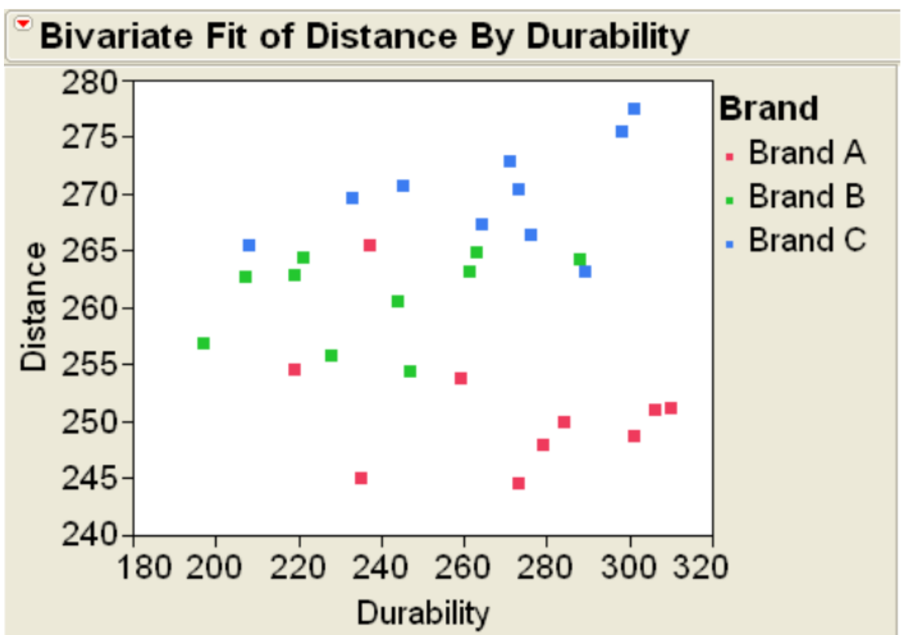
This is a **very strong** relationship. The daily amount of gas consumed can be predicted quite accurately for a given temperature value.

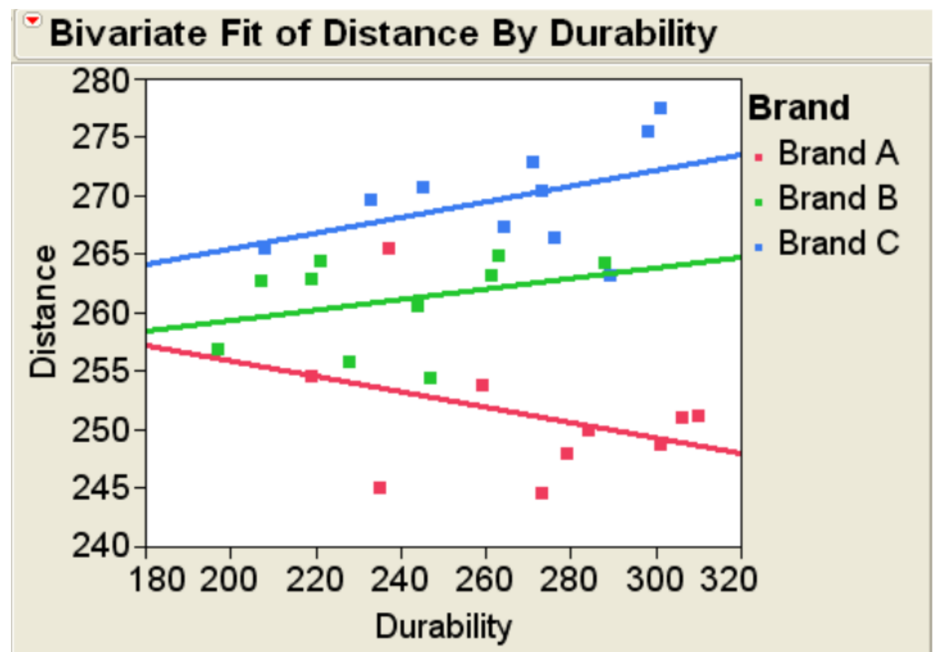
Multivariate relationships

Involving more than 2 variables

One example – colouring a scatterplot with a third (categorical) variable

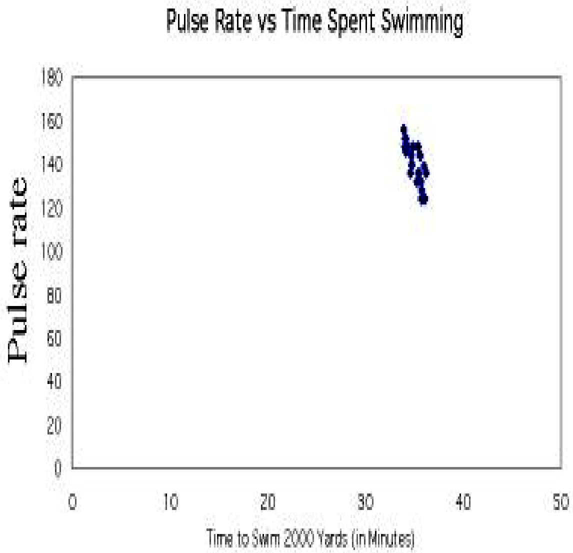
In Class Exercise 3.
What can you see from this graph?

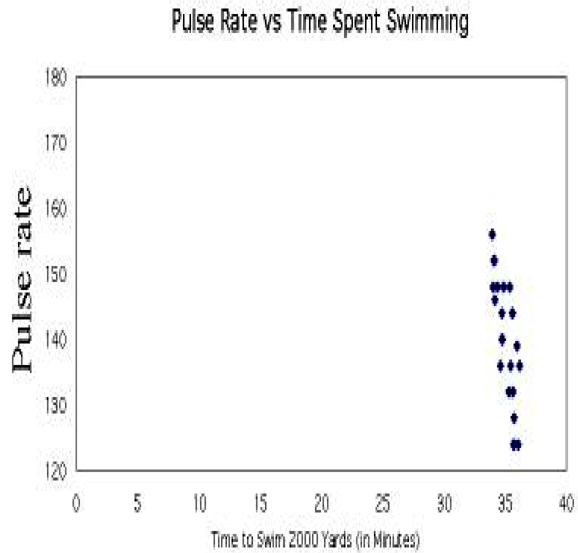


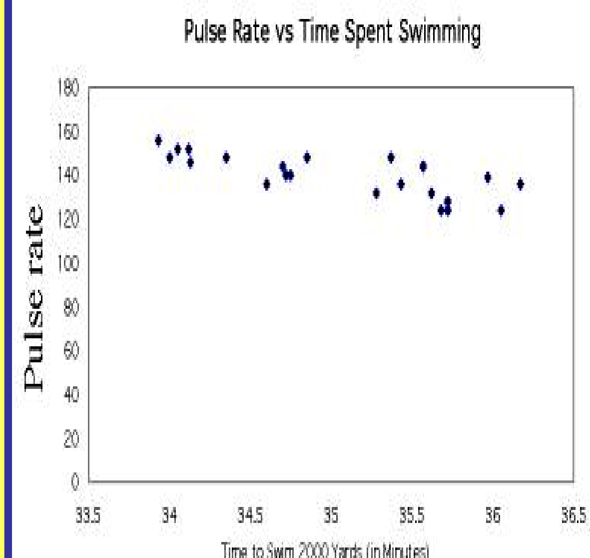


How to scale a scatterplot

Same data in all four plots







Using an inappropriate scale for a scatterplot can give an incorrect impression.

Both variables should be given a similar amount of space:

- Plot roughly square
- Points should occupy all the plot space (no blank space)

Aim 2.2 Numerically - The Pearson Sample Correlation coefficient *(r)*

- Measures the direction and strength of the linear relationship between two numerical variables.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{s_{xy}}{s_x s_y}$$

$\bar{x}$ be the sample mean of X;
$\bar{y}$ be the sample mean of Y;
s_x is the sample standard deviation of X;
s_y is the sample standard deviation of Y;
s_{xy} is the sample covariance between X and Y;
n be the number of observations

- R is used to compute this value.
- However note that the formula considers the variation in the *x* variable, in relation to the variation in the *y* variable).

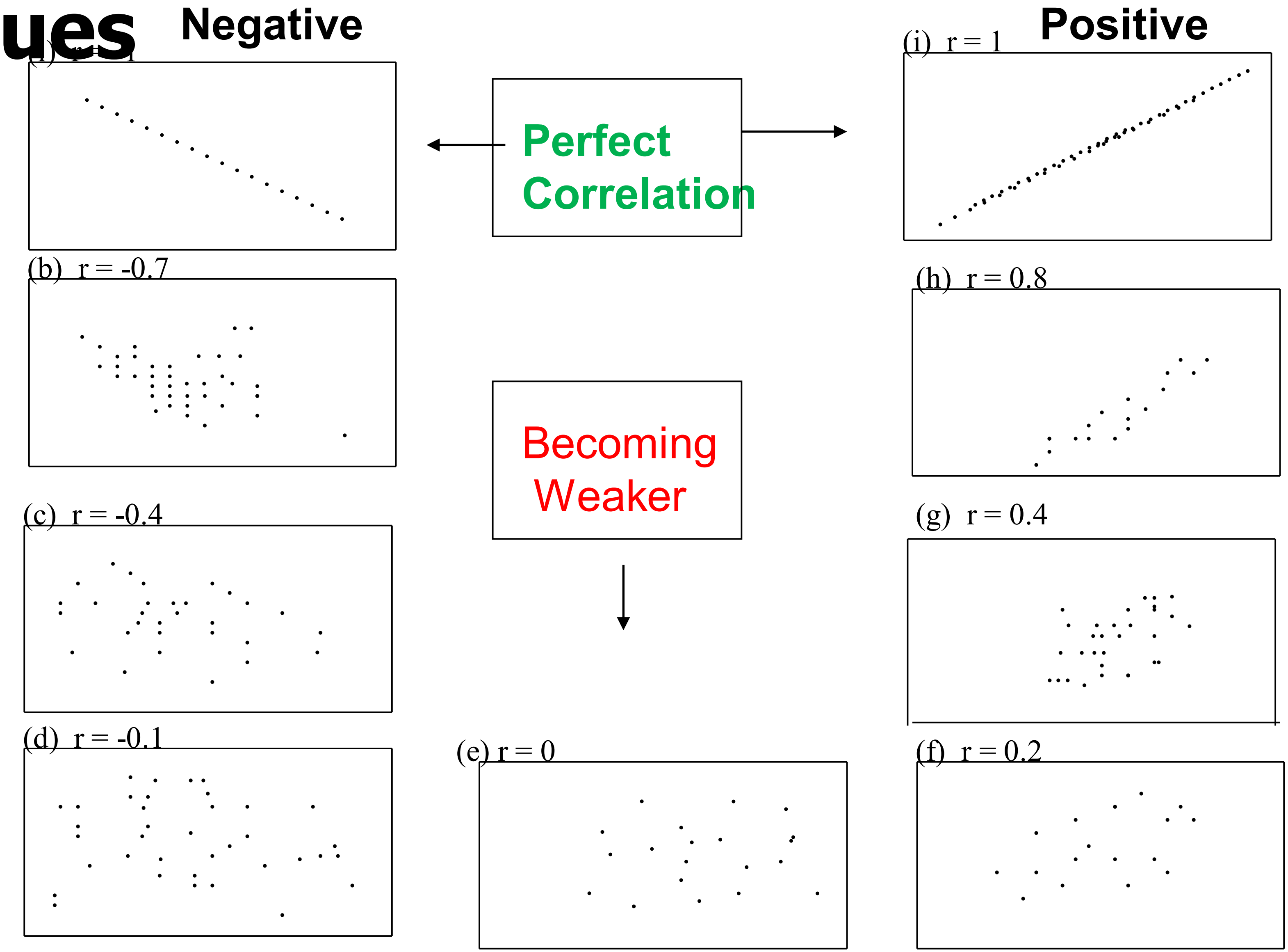
Understanding *correlation*

- Positive *(r)* indicates positive association between the variables

- Negative *(r)* indicates negative association between the variables.

- The correlation *(r)* always falls between -1 and +1.

Examples of the Pearson correlation values

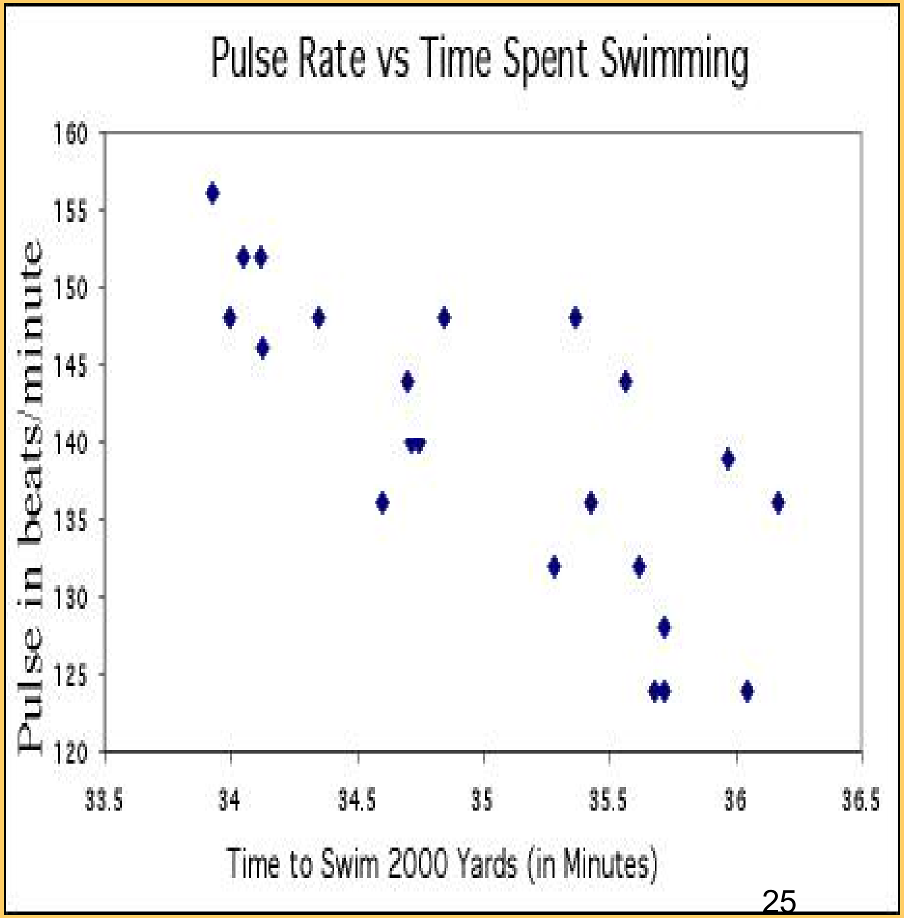


Example 5 The Pearson correlation coefficient *"r"*

Time to swim: $\bar{x}$= 35, s_x = 0.7

Pulse rate: $\bar{y}$ = 140 s_y = 9.5

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \frac{s_{xy}}{s_x s_y}$$



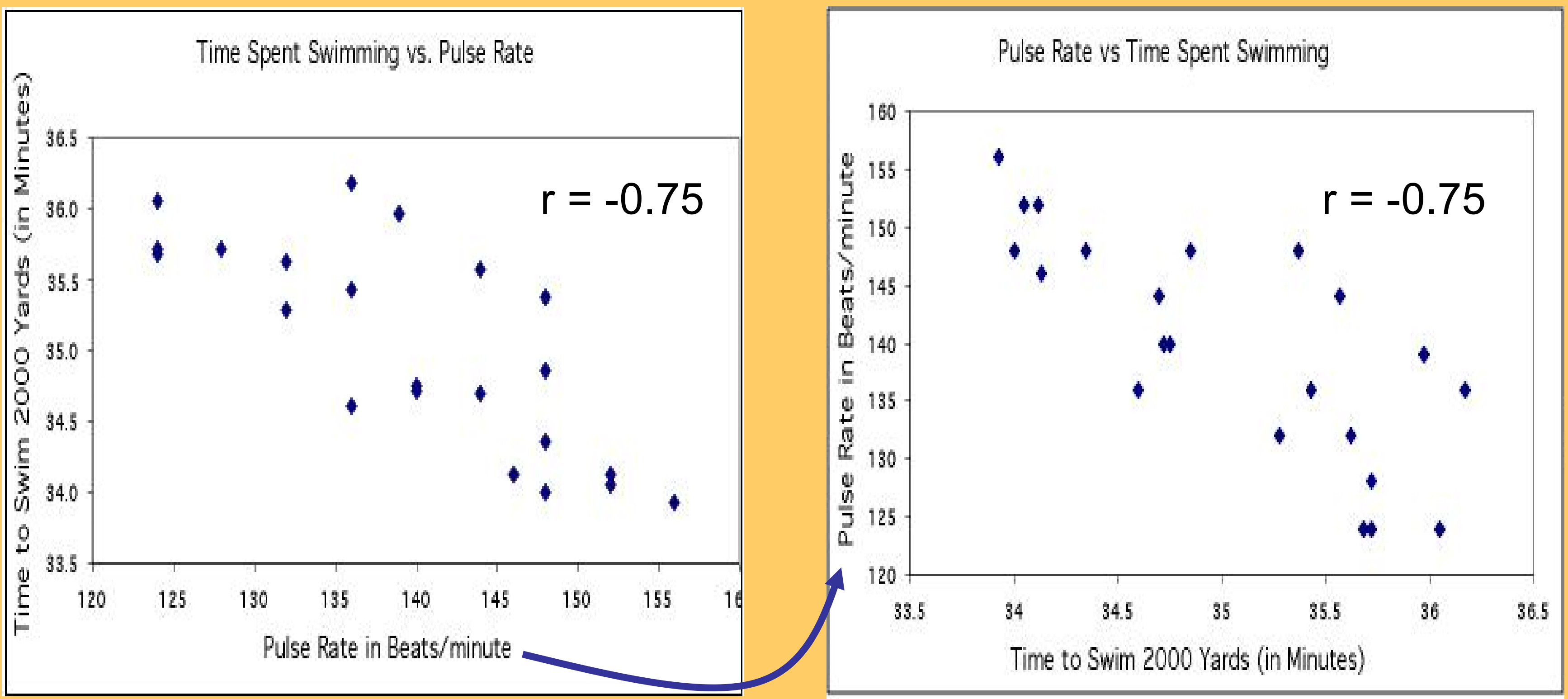
where
$\bar{x}$ be the sample mean of X
$\bar{y}$ be the sample mean of Y
s_x is the sample standard deviation of X;
s_y is the sample standard deviation of Y;
s_{xy} is the sample covariance between X and Y;
n be the number of observations

R:cor(x, y = NULL, use = "everything", method = c("pearson", "kendall", "spearman"))

Basic Properties 1: "*r*" does not distinguish x (explanatory) & y (response)

The correlation coefficient, *r*, treats x and y symmetrically. [Topic Notes 3.3]

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \frac{s_{xy}}{s_x s_y}$$



"Time to swim" is the explanatory variable here, and belongs on the *x* axis. However, in either plot *r* is the same (*r* = -0.75).

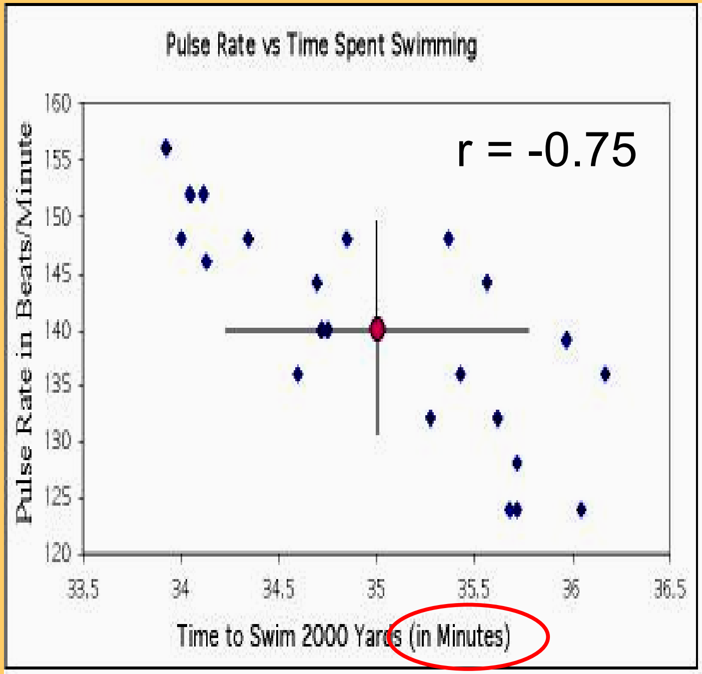
Basic Properties 2:

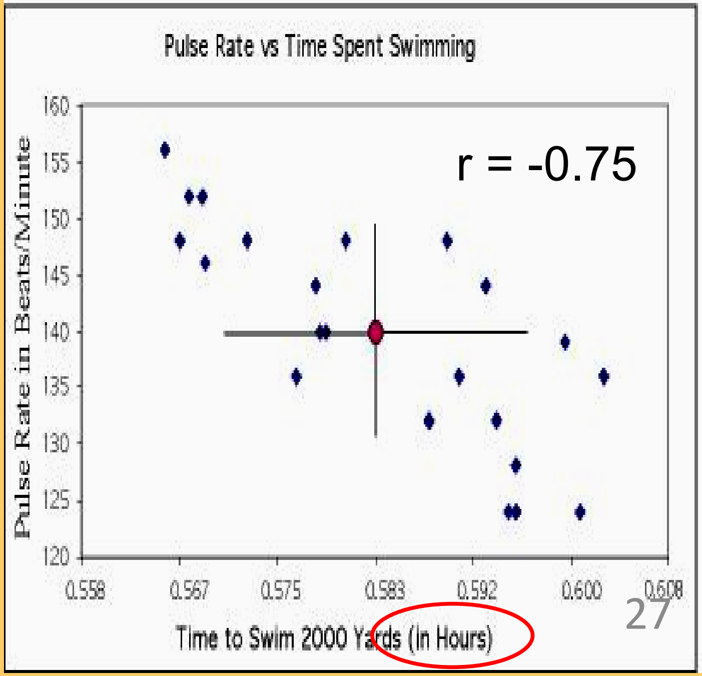
"*r*" has no unit

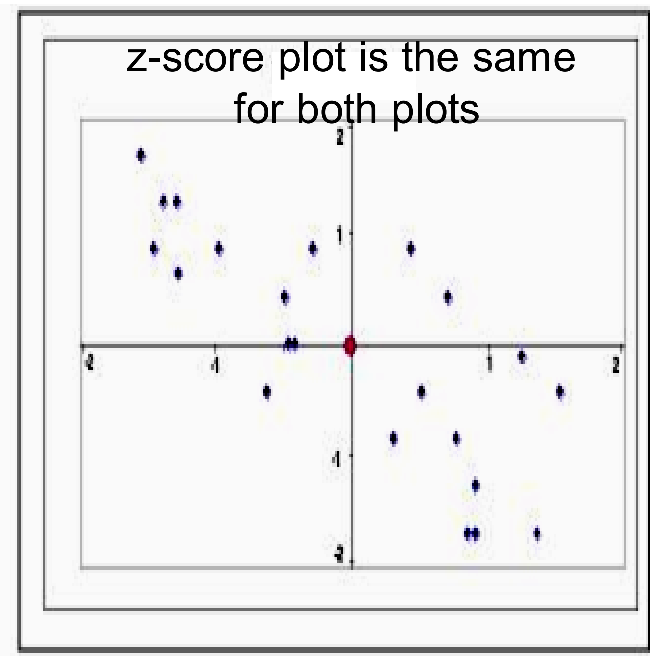
Changing the units of variables does not change the correlation coefficient "*r*", because we get rid of all our units when we standardize (get z-scores).

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

z for time *z for pulse*







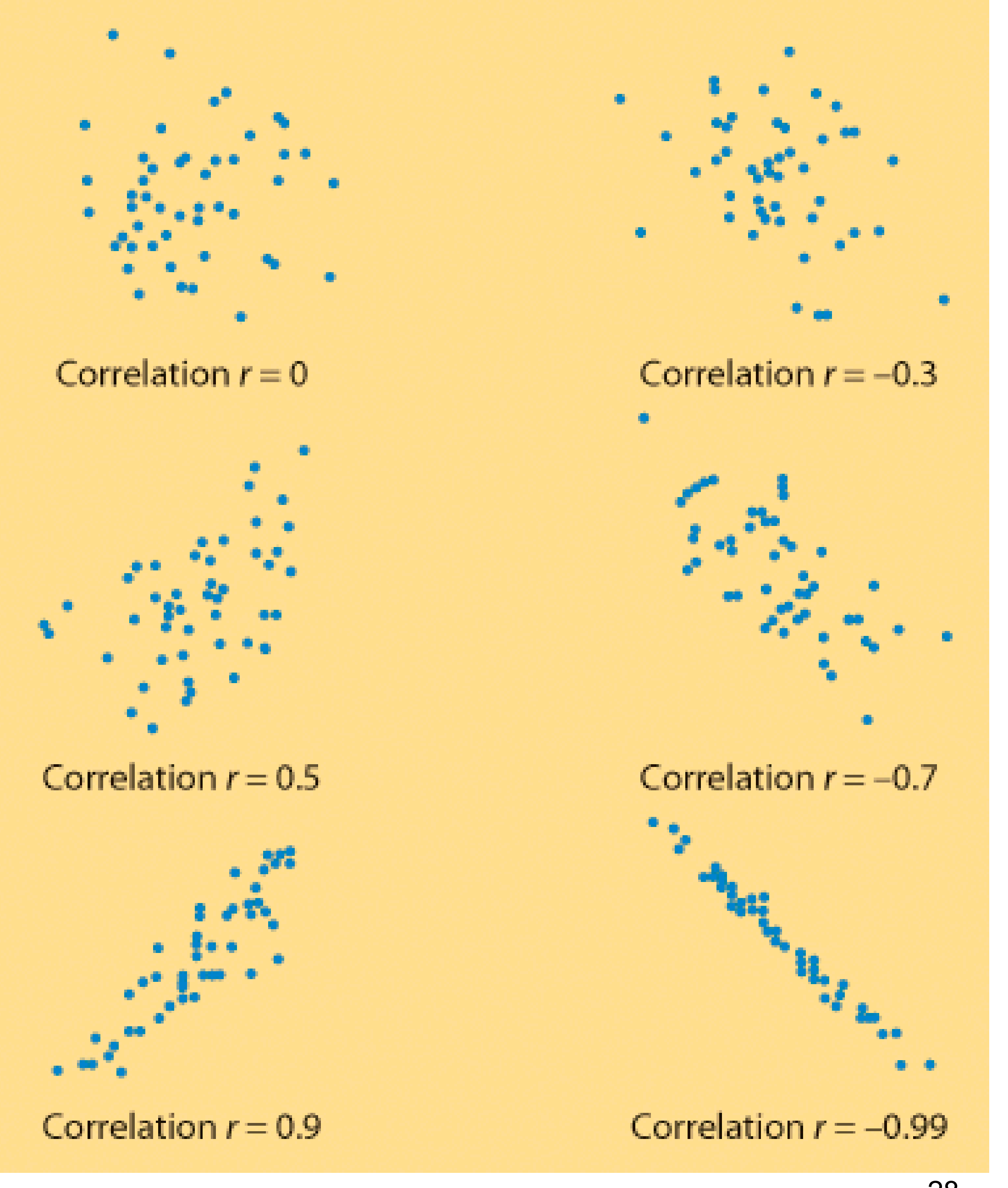
Basic Properties 3:

“r” ranges from -1 to +1

"*r*" quantifies the **strength** and **direction** of a linear relationship between 2 quantitative variables.

Strength: *how closely the points follow a straight line.*

Direction*: is positive when individuals with higher X values tend to have higher values of Y.*



Aim 2.3 Hypothesis Testing for a Linear Relationship

- The test of significance for ***ρ* (population correlation)** uses the one-sample *t*-test for ***H_0*: *ρ* = 0.**
- We compute the *t* statistic for sample size *n* and sample correlation coefficient *r*.

STEP 1 Hypotheses

H_0: $\rho = 0$ (no correlation)

H_1: $\rho \neq 0$ (correlation)

STEP 2 Test statistic $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

STEP 3 The sampling distribution

t ~ T with df (n-2)

$$t = \frac{r-\rho}{\sqrt{\dfrac{1-r^2}{n-2}}} \quad \text{where}$$

$$r = \sqrt{r^2} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

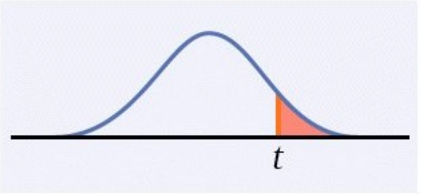
R:cor.test(x, y, alternative = c("two.sided", "less", "greater"), method = c("pearson", "kendall", "spearman"), exact = NULL, conf.level = 0.95, continuity = FALSE, ...)

Hypothesis Testing for a Linear Relationship

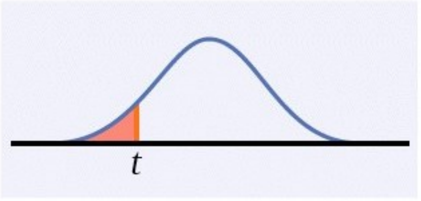
STEP 4. The P-Value

The P-value is the area under the sampling distribution $T(n-2)$ for values of T as or more extreme than t in the direction of H_a.

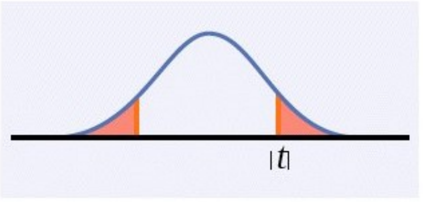
H_a: $\rho > 0$ is $P(T \geq t)$



H_a: $\rho < 0$ is $P(T \leq t)$



H_a: $\rho \neq 0$ is $2P(T \geq |t|)$



STEP 5 Decision

STEP 6 Conclusion

Inference for Correlation

- When the hypothesis H_0: $\rho = 0$ is rejected, it is safe to assume that there is some sort of relationship between the variables x and y.

- Recall that correlation measures ***linear*** relationships. As such, it is not always a reliable indicator of ***nonlinear*** relationships.

- When the hypothesis H_0: $\rho = 0$ is *not* rejected, do not assume that the variables are unrelated.

  ✓ First of all, it is possible that a Type II error may have occurred!

  ✓ Second, it is possible that x and y are related in a nonlinear way that the correlation coefficient r has no chance of detecting.

  ✓ A good way to investigate the second possibility is to examine a residual plot.

Assumptions underlying Pearson's correlation coefficient

- Both variables on equal interval/ratio scales
- Linear relationship between the variables
- Each variable has a normal distribution

Assumptions underlying Spearman's and Kendall's correlation coefficients

Independence assumption:

- $\{(X_i, Y_i)\}_{i=1}^{n}$ are iid from some bivariate population

Continuity assumption:

- $F_{X,Y}$ is a continuous distribution

Aim 3 Other Types of Correlations

- **Spearman's Correlation Coefficient(r_s)**
  - Correlation coefficients based on ranks
  - Ordinal data &/or non-normal distribution

- **Kendall**
  - To deal with data samples with tied ranks.
  - It is known as the Kendall's tau-b coefficient and is more effective in determining whether two non-parametric data samples with ties are correlated.
  - Nominal data

Which correlation?

- Number of minutes computer use (numerical) and level of discomfort (0-10) (ordinal close to discrete/numerical)
  – Pearson's
- Balance (good, moderate and poor) (ordinal) and number of days of treatment (numerical)
  – Spearman's
- Two interviewers ranked 10 candidates for a position (both ordinal).

  - Kendall's

From Pearson to Spearman

- When assumptions for Pearson's cannot be met then use Spearman's rank correlation coefficient.

- Pearson's
  - ✓ Measure only the degree of linear association
  - ✓ Based on the assumption of bivariate normality of two variables

- Spearman's
  - ✓ Take in account only the ranks
  - ✓ Measure the degree of monotone association
  - ✓ Inferences on the rank correlation coefficients are distribution-free

Aim 3.1 From Pearson to Spearman

Pearson

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{s_{xy}}{s_x s_y}$$

where

be the sample mean of X

be the sample mean of Y

s_x is the sample standard deviation of X;

s_y is the sample standard deviation of Y;

s_{xy} is the sample covariance between X and Y;

n be the number of observations

Spearman's Rank Correlation Coefficient

- $$r_s = \frac{\sum_{i=1}^{n}(u_i-\bar{u})(v_i-\bar{v})}{\sqrt{\{\sum_{i=1}^{n}(u_i-\bar{u})^2\}\{\sum_{i=1}^{n}(v_i-\bar{v})^2\}}} \quad (1)$$

  - Remark:
    - $u_i = rank(x_i)\ v_i = rank(y_i)$
    - $d_i = u_i - v_i$ are the difference in ranks
    - n=number of pairs of X's and Y's.

Spearman’s Rank Correlation Coefficient

- Steps for calculating r_s
  - Assign ranks to x_i’s and y_i’s . In case of ties, assign midranks.
  - Let u_i= rank (x_i), Let v_i= rank (y_i)
  - If u_i and v_i are integers, then a more convenient formula for calculating r_s:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where $d_i = u_i - v_i$ are the differences in ranks

- Like Pearson’s, Spearman’s correlation ranges from –1 to 1

Example 6a Act Ed Notes Ch 11 page 498

Calculate Spearman's rank correlation coefficient for the claims settlement data and comment.

Claim (£100's) x	2.10	2.40	2.50	3.20	3.60	3.80	4.10	4.20	4.50	5.00
Payment (£100's) y	2.18	2.06	2.54	2.61	3.67	3.25	4.02	3.71	4.38	4.45

Solution

For the claims settlement data:

Claim x	Payment y	Rank x	Rank y	d	d^2
2.1	2.18	1	2	–1	1
2.4	2.06	2	1	1	1
2.5	2.54	3	3	0	0
3.2	2.61	4	4	0	0
3.6	3.67	5	6	–1	1
3.8	3.25	6	5	1	1
4.1	4.02	7	8	–1	1
4.2	3.71	8	7	1	1
4.5	4.38	9	9	0	0
5	4.45	10	10	0	0

This gives:

$$r_s = 1 - \frac{6 \times 6}{10 \times (10^2 - 1)} = 0.9636$$

Large sample approximation: r_s

- For a large sample *n*,

$$E(r_s) = 0, \quad Var(r_s) = \frac{1}{n-1}$$

- Normal approximation of Spearman correlation is

$$Z = \frac{r_s - E(r_s)}{SD(r_s)} = \frac{r_s - 0}{\frac{1}{sqrt(n-1)}} = r_s\sqrt{(n-1)} \sim N(0,1)$$

Hypothesis Testing for ρ_s (Spearman population correlation)

- The test of significance for ρ_s uses approximate Normal distribution for H_0: $\rho_s = 0$.
- We compute the z statistic for sample size n and sample correlation coefficient r_s.

STEP 1 Hypotheses

H_0: X and Y are independent H_1: X and Y are dependent

STEP 2 Test statistic r_s (exact) or z=$r_s\sqrt{(n-1)}$ (approximate)

STEP 3 The sampling distribution based on ranks (exact) or z ~ N(0,1) (approximate)

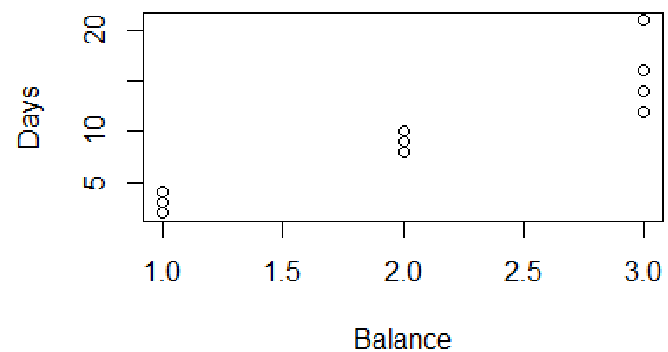
STEP 4 The p-value (see R output)

STEPS 5 and 6 Decision and Conclusion.

Example 6b

- Below is an example in which Balance is an ordinal data, Days is numerical.
- The histogram of Days is skewed to the right. Don't use Pearson.
- The sample correlation coefficient $r_s = 0.9494$ (see R output)

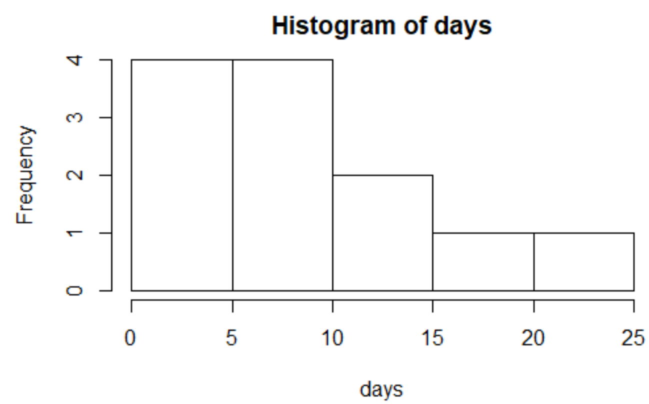
Balance	3	3	3	3	2	2	2	2	1	1	1	1
Days	12	14	16	21	10	9	8	10	4	2	3	2
	3 Good; 2 moderate; 1 poor											



R code

```
balance <-c(3,3,3,3,2,2,2,2,1,1,1,1)
days <- c(12,14,16,21,10,9,8,10,4,2,3,2)
cor(balance,days, method="spearman")
```

0.9494253



Example 7 Hypothesis Testing

Balance	3	3	3	3	2	2	2	2	1	1	1	1
Days	12	14	16	21	10	9	8	10	4	2	3	2
	3 Good; 2 moderate; 1 poor											

For Spearman's test, p-values are computed using algorithm AS 89 for n < 1290 and exact = TRUE, otherwise via the asymptotic t approximation. Note that these are 'exact' for n < 10, and use an Edgeworth series approximation for larger sample sizes (the cutoff has been changed from the original paper).

STEP 1 Hypotheses

H_0: X and Y are independent

H_1: X and Y are correlated

STEP 2 Test statistic

S=14.464 (see R output)

STEP 3 The sampling distribution (based on ranks)

STEP 4 The p-value = 2.393 x 10^(-6)

(using t approx, not Normal approx, see ?cor.test)

STEP 5 Decision: reject Ho

STEP 6 Conclusion: There is a correlation between Balance and Days.

R output

```
cor.test(balance,days,method="spearman", exact=F)
```

Spearman's rank correlation rho
data: balance and days
S = 14.464, p-value = 2.393e-06
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.9494253

Kendall's Rank Correlation

- Suppose we have n pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ where (X_i, Y_i) is i-th subject's data

- We want to make inferences about association between X and Y
  - Let $F_{X,Y}$ denote joint distribution of X and Y
  - Let F_X and F_Y denote marginal distributions of X and Y

- Null hypothesis is statistical independence: $F_{X,Y}(x, y) = F_X(x)F_Y(y)$ for all (x, y)

Kendall's Rank Correlation

- Parameter of interest is Kendall's Population correlation coefficient:

$$\tau = 2P[(Y_2 - Y_1)(X_2 - X_1) > 0] - 1$$

- The null hypothesis about τ is independence

$$H_0 : \tau = 0$$

and we could have one of three alternative hypotheses:

- One-Sided Upper-Tail: H_1: $\tau > 0$ (positively correlated)
- One-Sided Lower-Tail: H_1: $\tau < 0$ (negatively correlated)
- Two-Sided: H_1: $\tau \neq 0$ (correlated)

Kendall's Rank Correlation

- For all $n(n-1)/2$ pairs of observations (X_i, Y_i) and (X_j, Y_j) with $1 \leq i < j \leq n$, calculate paired sign statistic $Q[(X_i, Y_i), (X_j, Y_j)]$ where

$$Q[(a,b),(c,d)] = \begin{cases} 1 & \text{if } (d-b)(c-a) > 0 \\ -1 & \text{if } (d-b)(c-a) < 0 \end{cases}$$

- **The Kendall test statistic K** is defined as $$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q[(X_i, Y_i), (X_j, Y_j)]$$

  which is simply the sum of the paired sign statistic for all pairs.

Can estimate population τ using sample estimate

$$\hat{\tau} = \frac{2K}{n(n-1)} = \bar{K}$$

given that $-\frac{n(n-1)}{2} \leq K \leq \frac{n(n-1)}{2}$.

$\hat{\tau}$ is sometimes referred to as Kendall's τ rank correlation coefficient.

Large sample approximation for the test statistic *K*

R code
```
cor(x,y, method="kendall")
cor.test(x,y, method="kendall")
```

- For a large sample *n*,

$$E(K) = 0, \quad Var(K) = \frac{n(n-1)(2n+5)}{18}$$

- We can create a standardized test statistic K^* of the form

$$K^* = \frac{K - E(K)}{\sqrt{Var(K)}}$$

which asymptotically follows a N(0, 1) distribution.

- We will not calculate the correlations manually. We will use R output.

Example 8 Test Scores of Male Twins

Nonparametric Statistical Methods, 3rd Ed. (Hollander et al., 2014)

Table 8.5 Psychological Test Scores of Dizygous Male Twins

Pair i	X_i	Y_i
1	277	256
2	169	118
3	157	137
4	139	144
5	108	146
6	213	221
7	232	184
8	229	188
9	114	97
10	232	231
11	161	114
12	149	187
13	128	230

Source: P.J. Clark, S. G. Vandenberg, and C. H. Proctor (1961).

R output

```
x=c(277,169,157,139,108,213,232,229,114,232,161,149
,128)
y=c(256,118,137,144,146,221,184,188,97,231,114,187,
230)

cor.test(x,y,method="kendall",alternative="greater")
Kendall's rank correlation tau  data: x and y
z = 1.6503, p-value = 0.04944
alternative hypothesis: true tau is greater than 0
sample estimates:
tau
0.3483943
```

STEP 1 Hypotheses

H_0: X and Y are independent

H_1: X and Y are positively correlated

STEP 2 Test statistic

Z=1.6503 (see R output)

STEP 3 The sampling distribution

STEP 4 The p-value = 0.04944

STEP 5 Decision: reject Ho

STEP 6 Conclusion: There is a correlation between X and Y

R:

For Kendall's test, by default (if exact is NULL), an exact p-value is computed if there are less than 50 paired samples containing finite values and there are no ties. Otherwise, the test statistic is the estimate scaled to zero mean and unit variance, and is approximately normally distributed.