

17-200s  $\left\{ \begin{array}{l} 7-10\text{Qs (Ch 11)} \\ 7-11\text{Qs (Ch 12)} \end{array} \right.$

## STATS 10X WORKSHOP

**SATURDAY V: CHAPTERS 11 & 12**

**SATURDAY 23 OCTOBER 2010**



Students **MUST REGISTER** for all workshops with  
**The Student Learning Centre, 3<sup>rd</sup> Floor, Information Commons**



## Statistical help available at the SLC

The Student Learning Centre (SLC) offers help for STATS 10x by offering:

- one-on-one tutoring help, and
- a number of workshops

### ***One-on-one help over S2 2010 including exam period***

One-on-one assistance for STATS 10x is available at the SLC. Check appointment availability and book at SLC reception in person (go to the third floor of the Information Commons building) or by calling 373-7599 ext. 88850.

### ***SLC STATS 10x Exam Prep Workshops***

Any questions regarding STATS 10x workshops should be forwarded to:

**Leila Boyle**; SLC Statistics Co-ordinator: [l.boyle@auckland.ac.nz](mailto:l.boyle@auckland.ac.nz)

These twelve workshops (six different sessions, each repeated twice) are held prior to the exam, from Saturday 2 October until Monday 1 November 2010 (inclusive).

These workshops concentrate on questions reviewing the **basic concepts**, rather than questions on finer details. They are designed to assist students to achieve a pass and **don't cover all material**.

**The timetable for these workshops is available with this handout. Currently the SLC website is still partly down so online enrolments are not available until further notice. Please enrol in each of your preferred classes at the SLC by:**

- **Going to SLC reception in person**
- **Emailing [slc@auckland.ac.nz](mailto:slc@auckland.ac.nz) with your name, ID number and the workshop/s you wish to attend.**
- **Calling SLC reception on 373-7599 ext. 88850 to enrol over the phone. Make sure you know which workshop/s you want to enrol in and have your ID number handy.**

### **Useful Websites**

- SLC webpage: [www.slc.auckland.ac.nz](http://www.slc.auckland.ac.nz) (The SLC website currently has all functionality except online enrolment! The undergraduate brochure with information on all upcoming workshops is available for download here).
- Cecil: <https://cecil.auckland.ac.nz>
- Leila's website for STATS 101/108 SLC workshop handouts & information: [www.stat.auckland.ac.nz/~leila](http://www.stat.auckland.ac.nz/~leila)

1 or 2  
qual. var.s

# Revision Notes

7-100s

## Chapter 11 Review – Chi-Square Tests

Look at blue pages for good notes and test/exam questions for practice

- We use **Chi-Square** tests to test **proportions** from **tables of counts**.
- There are 2 kinds of Chi-Square tests:
  1. For **one-way tables of counts**: Test for **goodness of fit**
  2. For **two-way tables of counts**: Test for **independence**
- We determine which Chi-Square test to use by the number of samples taken and the number of qualitative variables / factors to be tested.

1 qual var.

2 qual. var.s

- **Assumptions** for Chi-Square tests to be **valid** are:

- **At least 80%** of the **expected counts** must be **5 or more**, and
- **Every expected count** must be **greater than 1**.

- The **test-statistic** for Chi-Square tests is the **sum** of all the cell contributions from the table:

$$\chi_0^2 = \sum_{\text{all cells in the table}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum \frac{(O - E)^2}{E}$$

on formulae sheet

### 1. One-Way Tables of Counts – Chi-square test for Goodness of Fit

- Use when you have **1 sample** & **1 factor** of interest.
- **Hypotheses**

1 qual. var.

$H_0$ : The data come from the specified distribution.

$H_1$ : The data do not come from the specified distribution.

- **Expected Count:**

For one-way tables:  
Expected count in  $j^{\text{th}}$  cell =  $p_j \times n$   
where  $p_j$  is the specified cell's probability

- **Chi-square test-statistic:**

$$\chi_0^2 = \sum_{\text{all cells in the table}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

- **Degrees of freedom**

For one-way tables:  
 $df = J - 1$

where  $J$  is the number of categories.

# Cross-classify

## 2. Two-Way Tables of Counts - Chi-square test for Independence

- Use when you have **1 sample & 2 factors** of interest.

OR **2 or more independent random samples & 1 factor** of interest

- Hypotheses**

$H_0$ : the two factors are **independent** / not associated / not related

$H_1$ : the two factors are **not independent** / dependent / associated / related

- Expected Count**

For two-way tables:  
Expected count in cell  $(i, j) = \frac{R_i C_j}{n}$

where  $R_i$  is the row total  
and  $C_j$  is the column total  
and  $n$  is the table total

- Chi-square test-statistic:**

$$\chi^2_0 = \sum_{\text{all cells in the table}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

- Degrees of freedom**

For two-way tables:  
 $df = (I - 1)(J - 1)$

where  $I$  is the number of rows  
and  $J$  is the number of columns

same family or distribution

### Another way of looking at it - Chi-square test for Homogeneity

The null hypothesis is often written as a statement of **homogeneity** (sameness).

$H_0$ : the underlying distribution of variable 1 is **the same** for each level of variable 2.

$H_1$ : the underlying distribution of variable 1 is **not the same** at all levels of variable 2.

The sampling situation determines which one of the two variables is variable 1 and which one is variable 2. There are two possibilities:

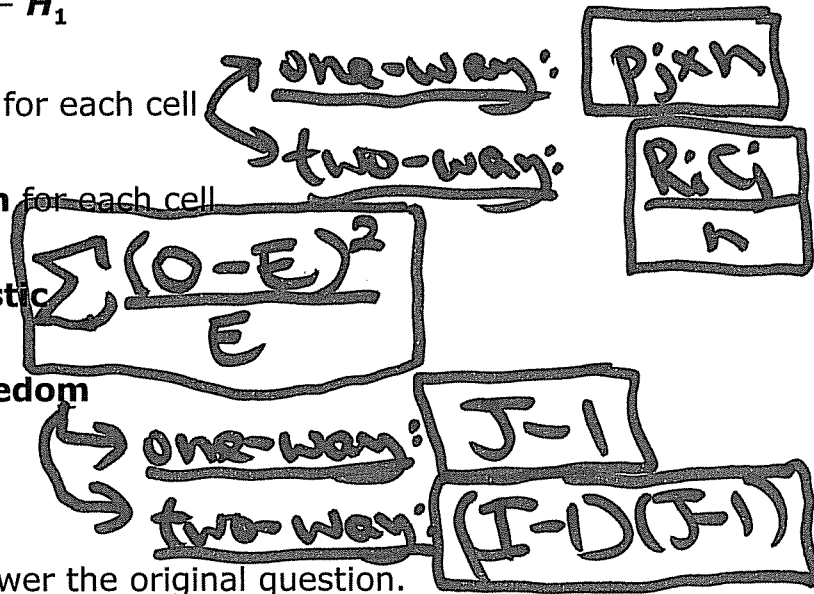
- If you have **2 or more independent samples** taken from **different populations** (variable 2) and **each sample** is then **divided** up by a **factor/qualitative variable** (variable 1) then the null hypothesis can be a statement of homogeneity among the populations from which the samples have been taken:

$H_0$ : The distribution of variable 1 is the same for each population (variable 2)

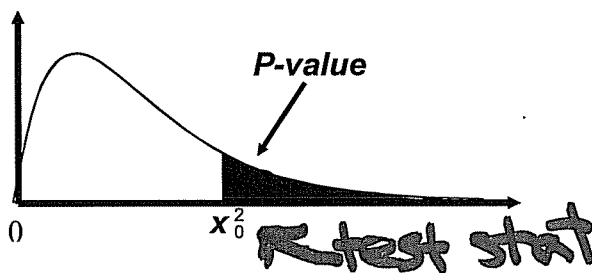
- If a **single random sample** has been **cross classified** by variable 1 and variable 2 then the null hypothesis can be either a statement of homogeneity (sameness) on the **rows** or on the **columns**, i.e. it doesn't matter which variable is variable 1 and which variable is variable 2, the hypotheses above are completely interchangeable.

## Steps in performing a **Chi-square test**

1. Identify which **situation** you have (either 1 sample or 2 or more independent samples)
2. State the null hypothesis –  $H_0$
3. State the alternate hypothesis –  $H_1$
4. Calculate the **expected count** for each cell
5. Calculate the **cell contribution** for each cell
6. Find the Chi-square **test statistic**
7. Determine the **degrees of freedom**
- ~~8. Find the **P-value**.~~
9. **Interpret** the  $P$ -value and answer the original question.



### **P-value**



- $P\text{-value} = \text{pr}(X^2 \geq x_0^2)$ , where  $X^2 \sim \text{Chi-square}(df)$

The **bigger** the test statistic is, the **stronger** the evidence **against**  $H_0$ .

- As with the  $t$ -test and the  $F$ -test, the  $P$ -value is the conditional probability of observing a test statistic as extreme as that observed or more so, **given that** the null hypothesis is true.
- If you have a **small**  $p$ -value – note which **cells contributed the most** to the **large** Chi-square statistic, i.e. look for cells where there are **large differences** between the **observed** counts and the **expected** counts.
- Look at the size of the differences: create confidence intervals for differences between two proportions (work out the correct standard error formula).

## Chapter 11 – Questions

1. Which one of the following tools give the best display of the relationship between 2 qualitative variables:

- (1) Scatterplot
- (2) Boxplot
- (3) Dotplot
- (4) Table of counts
- (5) One-way analysis of variance

2. Which **one** of the following statements is **false**:

T (1) A Chi-square test for goodness of fit is used to carry out a formal analysis on data presented in a one-way table of counts.

T (2) If one or more of the expected counts in a table is less than 1 then we would have concerns with the validity of a Chi-square test carried out on these data.

T (3) If, for several cells in a table of counts, there are relatively large differences between the observed counts and the expected counts under the null hypothesis, then the *P-value* for a Chi-square test will be small.

(4) The greater the value of the Chi-square test statistic, the ~~weaker~~ **stronger** the evidence against the null hypothesis.

T (5) The Chi-square test statistic is a measure of the difference, over all cells in the table, between the counts observed from the sample and the counts that would have been expected under the null hypothesis.

3. Which **one** of the following statements is **true**:

F (1) If, for all cells in a table of counts, there are relatively small differences between the observed counts and the expected counts under the null hypothesis, then the data provides ~~evidence~~ **no** evidence against the null hypothesis.

F (2) The greater the value of the Chi-square test statistic, the ~~larger~~ **smaller** the *P-value*.

F (3) For a Chi-square test to be valid the total count in the table,  $n$ , is required to be ~~small~~ **relatively large**.

F (4) The *P-value* in a Chi-square test is the probability, given that the null hypothesis is true, of obtaining a test statistic as extreme, or ~~less so~~ **more** as that observed.

(5) If the *P-value* is small, then the cells with the largest contributions to the test statistic show which cells have observed counts that are far different (relatively) from those expected under the null hypothesis.



7. The expected count for each of the three mayoral candidates is:

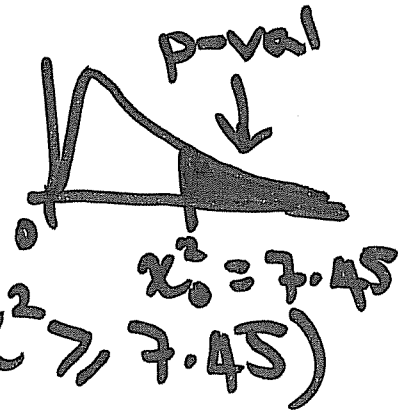
- (1) 39.6, 32.3, 27.9
- (2) 100, 100, 100
- (3) 27.9, 32.3, 39.6
- (4) 90.9, 90.9, 90.9
- (5) 40.0, 40.0, 40.0

$$p_j \times n = \frac{1}{3} \times 300 = 100$$

8. The Chi-square test statistic is 7.45, to two decimal places. The *P*-value for this test is calculated by:

- (1)  $\Pr(X^2 \leq 7.45)$
- (2)  $2 \times \Pr(X^2 \geq 7.45)$
- (3)  $\Pr(X^2 \geq 7.45)$
- (4)  $2 \times \Pr(X^2 < 7.45)$
- (5)  $\Pr(0 \leq X^2 \leq 7.45)$

- where  $X^2 \sim \text{Chi-square}(df)$
- where  $X^2 \sim \text{Chi-square}(df)$
- where  $X^2 \sim \text{Chi-square}(df)$
- where  $X^2 \sim \text{Chi-square}(df)$
- where  $X^2 \sim \text{Chi-square}(df)$



9. Which **one** of the following statements is **false**:

- (1) A Chi-square test for independence is used to carry out formal analyses on data presented in two-way tables of counts.
- (2) The Chi-square test statistic is a measure of the difference between what we see in the data and what we would expect to see if the null hypothesis was true.
- (3) If all the expected counts in a table are less than 5 then we would have no concerns with the validity of a Chi-square test carried out on these data.
- (4) If, for one or more cells in a table of counts, there are relatively large differences between the observed counts and the expected counts under the null hypothesis, then the data provides evidence against the null hypothesis.
- (5) The *P-value* in a Chi-square test is the probability, assuming that the null hypothesis is true, of observing data at least as discrepant as that obtained in the data on which the test is carried out.

Questions 13 to 17 are about the following information.

Since 25 August 1997, David has been recording the number of junk e-mails he receives each day. Part of his recorded data is listed in the table below.

Number of Junk e-mails Received

Week Beginning	Mon	Tue	Wed	Thu	Fri	Total
25 Aug 97	8	1	1	2	1	13
1 Sep 97	15	1	3	0	3	22
8 Sep 97	5	1	4	2	5	17
.	.	.	.	.	.	.
.	.	.	.	.	.	.
2 Feb 98	6	2	1	1	0	10
9 Feb 98	4	1	1	1	1	8
<b>Total</b>	<b>160</b>	<b>64</b>	<b>58</b>	<b>51</b>	<b>55</b>	<b>388</b>

3x  
388  
= 116.4

.2 x 388  
= 77.6

.1 x 388  
= 38.8

expected: 116.4, 77.6, 77.6, 77.6, 38.8, 388

David wanted to test the probability that the percentage of junk emails he received (for any given working week) were on Mondays 30%, Tuesdays 20%, Wednesdays 20%, Thursdays 20% and Fridays 10%.

13. David tested this probability based on the total number of emails he received over the 6-month period. From the table above, what are the expected counts for the total emails received on each day?

- (1) 77.6, 77.6, 77.6, 77.6, 77.6
- (2) 48, 12.8, 11.6, 10.2, 5.5
- (3) 38.8, 77.6, 77.6, 77.6, 116.4
- (4) 116.4, 77.6, 77.6, 77.6, 38.8
- (5) 3.9, 2.2, 3.4, 2, 0.8

H<sub>0</sub>: p<sub>M</sub> = .3; p<sub>Tu</sub> = p<sub>W</sub> = p<sub>Th</sub> = .2; p<sub>F</sub> = .1

H<sub>1</sub>: at least one p<sub>j</sub> is diff. from that specified

14. The value of the Chi-square statistic is:

- (1)  $\frac{(160-116.4)^2}{160} + \frac{(64-77.6)^2}{64} + \frac{(58-77.6)^2}{58} + \frac{(51-77.6)^2}{51} + \frac{(55-38.8)^2}{55}$
- (2)  $\frac{(160-116.4)^2}{116.4} + \frac{(64-77.6)^2}{77.6} + \frac{(58-77.6)^2}{77.6} + \frac{(51-77.6)^2}{77.6} + \frac{(55-38.8)^2}{38.8}$
- (3)  $\frac{(160-116.4)^2}{116.4} + \frac{(64-77.6)^2}{77.6} + \frac{(58-77.6)^2}{77.6} + \frac{(51-77.6)^2}{77.6} + \frac{(55-38.8)^2}{38.8}$
- (4)  $\frac{(64-77.6)^2}{77.6} + \frac{(58-77.6)^2}{77.6} + \frac{(51-77.6)^2}{77.6}$
- (5)  $\frac{(160-116.4)^2}{38.8} + \frac{(64-77.6)^2}{77.6} + \frac{(58-77.6)^2}{77.6} + \frac{(51-77.6)^2}{77.6} + \frac{(55-38.8)^2}{116.4}$

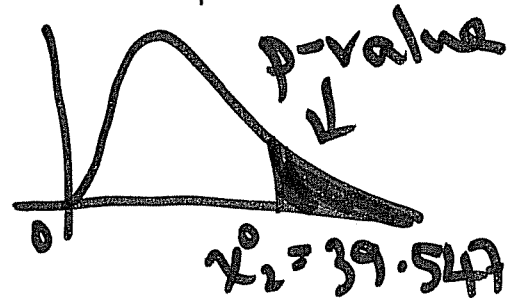
M.F?

$\chi^2$

15. The sum of the chi-square test statistic is 39.547. Which day during the 6-month period contributed the greatest to this test statistic?
- (1) Monday
  - (2) Tuesday
  - (3) Wednesday
  - (4) Thursday
  - (5) Friday

16. Which of the following statements is true for the above chi-square test?

- ~~(1)~~  $pr(\chi^2 \leq 39.547) = 1 - pr(\chi^2 \leq 39.547)$
- ~~(2)~~  $pr(\chi^2 \geq 39.547) = 1 - pr(\chi^2 \leq 39.547)$
- ~~(3)~~  $pr(\chi^2 \leq 39.547)$  not p-value
- (4)  $pr(\chi^2 \geq 39.547) = 1 - pr(\chi^2 \leq 39.547)$
- ~~(5)~~  $pr(\chi^2 = 39.547)$  not p-value



17. The P-value for  $H_0: p_1 = 0.30, p_2 = 0.20, p_3 = 0.20, p_4 = 0.20, p_5 = 0.10$  is 0.000. The best interpretation is?

- ~~(1)~~ The P-value of 0.000 provides no evidence against  $H_0: p_1 = 0.30, p_2 = 0.20, p_3 = 0.20, p_4 = 0.20, p_5 = 0.10$ .
- ~~(2)~~ The P-value of 0.000 provides weak evidence against  $H_0: p_1 = 0.30, p_2 = 0.20, p_3 = 0.20, p_4 = 0.20, p_5 = 0.10$ .
- ~~(3)~~ The P-value of 0.000 provides some evidence against  $H_0: p_1 = 0.30, p_2 = 0.20, p_3 = 0.20, p_4 = 0.20, p_5 = 0.10$ .
- (4) The P-value of 0.000 provides very strong evidence of  $H_0: p_1 = 0.30, p_2 = 0.20, p_3 = 0.20, p_4 = 0.20, p_5 = 0.10$ .
- (5) The P-value of 0.000 provides very strong evidence against  $H_0: p_1 = 0.30, p_2 = 0.20, p_3 = 0.20, p_4 = 0.20, p_5 = 0.10$ .

Questions 18 to 26 refer to the following information.

During 1999, students enrolled in stage one statistics at the University of Auckland were surveyed regarding their access to, and experience with, computers. The survey was included as a question in an assignment, and students were given marks for completing it (irrespective of the answers they gave). Staff administering the courses wished to use the results of this survey to draw conclusions about future stage one statistics students.

One question asked: *(cross-classified)* "At the start of the course, how would you describe your Excel experience?" A total of 918 students answered this question. Each of the 918 answers was classified according to the response given by the student, and the stream the student attended. The results are given in the table below, where 107, 108 and 101 refer to the various streams.

Response	Stream			Total
	107	108	101	
None	15	36	102	153
Very Little	44	89	119	252
Some	74	150	200	424
Lots	9	29	51	89
Total	142	304	472	918

*Q19 - Q24*

Table 4: Responses to question regarding Excel experience.

18. The variable Stream is:

- (1) discrete.
- (2) quantitative.
- (3) qualitative.
- (4) dependent.
- (5) continuous.

Questions 19 to 24 refer to the following additional information.

A staff member, before seeing the data, suggested that the true proportion of students falling in the categories None, Very Little, Some and Lots, irrespective of stream, were  $p_1 = 0.15$ ,  $p_2 = 0.30$ ,  $p_3 = 0.45$ ,  $p_4 = 0.10$  respectively. We wish now to test this hypothesis.

19. The most appropriate test to use is a:

- (1) Chi-square test of homogeneity.
- (2) two independent sample  $t$ -test.
- (3) Chi-square test of independence.
- (4) Chi-square test for goodness-of-fit.
- (5) one sample  $t$ -test.

20. Suppose it is appropriate to conduct a Chi-square test for goodness-of-fit (Note: This may not be correct). The null and alternative hypotheses for this test are:

~~(1)~~  $H_0: p_1 = 0.15, p_2 = 0.30, p_3 = 0.45, p_4 = 0.10$  ✓

~~$H_1: p_1 = 0.167, p_2 = 0.275, p_3 = 0.462, p_4 = 0.097$~~

~~(2)~~  $H_0: p_1 = p_2 = p_3 = p_4 = 0.25$  ✗

~~$H_1: \text{The probabilities in the null hypothesis are not all equal.}$~~

~~(3)~~  $H_0: n_1 = 153, n_2 = 252, n_3 = 424, n_4 = 89$  ✗

~~$H_1: \text{At least two of the totals in the null hypothesis are not correct.}$~~

~~(4)~~  $H_0: p_1 = 0.167, p_2 = 0.275, p_3 = 0.462, p_4 = 0.097$  ✗ *Dis!*

~~$H_1: \text{At least two of the probabilities in the null hypothesis are not correct.}$~~

(5)  $H_0: p_1 = 0.15, p_2 = 0.30, p_3 = 0.45, p_4 = 0.10$

$H_1: \text{At least two of the probabilities in the null hypothesis are not correct.}$

21. The Chi-square test for goodness-of-fit is conducted, and the results are given below. Each cell in the table gives the component of Chi-square (i.e. the cell contribution to the test statistic).

Response	None	Very Little	Some	Lots
Component of Chi-square	1.7		0.287	0.085

test statistic = 4.06

P-value = 0.255

Table 5: Results of Chi-square test.

Using Table 5, the correct value of c is:

(1) 0.275

(2) 1.99

(3) -1.07

(4) 1.07

(5) unknown.

$$4.06 - 1.7 - 0.287 - 0.085 = 1.988$$

$$\frac{(O-E)^2}{E}$$

22. The correct degrees of freedom for the test in Table 5 are:

(1) 2

(2) 12

(3) 4

(4) 6

(5) 3

$$df = J - 1 = 4 - 1 = 3$$

Questions 23 to 26 refer to the following additional information.

We now wish to conduct a Chi-square test on the full data set in Table 4 on page 13.

23. Which one of the following statements is true?

- ~~(1)~~ The data were collected as several samples with a single response factor. Therefore either a Chi-square test of homogeneity or a Chi-square test of independence is appropriate.
- ~~(2)~~ The data were collected as several samples with a single response factor. Therefore only a Chi-square test of homogeneity is appropriate.
- ~~(3)~~ The data were collected as one sample, cross-classified by two factors. Therefore only a Chi-square test of independence is appropriate. *or homogeneity*
- (4) The data were collected as one sample, cross-classified by two factors. Therefore either a Chi-square test of homogeneity or a Chi-square test of independence is appropriate. *or indep*
- ~~(5)~~ The data were collected as one sample, cross-classified by two factors. Therefore only a Chi-square test of homogeneity is appropriate. *or indep*

24. Suppose it is appropriate to conduct a Chi-square test of independence on these data. (Note: This may not be correct.) Based on the information in Table 4 (page 13), the estimated expected count associated with the (Lots, 108) cell is:

- (1) 29
- (2) 2.8
- (3) 0.01
- (4) 9.6
- (5) 29.5

$$\text{exp} = \frac{R_i C_j}{n} = \frac{89 \times 304}{918} = 29.5$$

25. The *P-value* for the test described in question 24 is 0.0014. Which one of the following statements gives the best interpretation of this *P-value*?

- (1) There is strong evidence that a student's response to the question 'At the start of the course, how would you describe your Excel experience?' is independent of the stream they are in.
- (2) There is strong evidence that a student's response to the question 'At the start of the course, how would you describe your Excel experience?' is not independent of the stream they are in.
- (3) There is ~~evidence~~ that a student's response to the question 'At the start of the course, how would you describe your Excel experience?' is very strongly related to the stream they are in.
- (4) There is ~~weak~~ evidence that a student's response to the question 'At the start of the course, how would you describe your Excel experience?' is related to the stream they are in.
- (5) There is ~~no~~ evidence that a student's response to the question 'At the start of the course, how would you describe your Excel experience?' is associated with the stream they are in.

Questions 26 refer to the following additional information.

Table 6 below contains some results from conducting the test described in question 24.

			Stream			
			107	108	101	Total
Response None	Count		15	36	102	153
	Expected Count		23.7	50.7	78.7	153.0
	Cell contribution		3.17	4.25	6.92	
Very Little	Count		44	89	119	252
	Expected Count		39.0	83.5	129.6	252.0
	Cell contribution		0.65	0.37	0.86	
Some	Count		74	150	200	424
	Expected Count		65.6	++	++	424.0
	Cell contribution		1.08	0.66	1.49	
Lots	Count		9	29	51	89
	Expected Count		13.8	++	++	89.0
	Cell contribution		1.65	++	0.60	
Total	Count		142	304	472	918
	Expected Count		142.0	304.0	472.0	918.0

Q24  
(= 29.5)

v. small

test stat,  $\chi^2_0$

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	21.697 <sup>a</sup>	6	0.0014
Likelihood Ratio	22.246	6	0.0011
Linear-by-Linear Association	3.827	1	0.0504
N of Valid Cases	918		

p-value

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 13.77.

Table 6: Chi-square test output.

$$df = (I-1)(J-1) = (4-1)(3-1) = 3 \times 2 = 6$$

26. Based on the information in Table 4 (page 13) and Table 6 above, which one of the following provides the best explanation for the small  $P$ -value (0.0014)?

- (1) There are fewer 107 and 108 students, and more 101 students in the None category, than would be expected under the assumption of independence.
- (2) There are around the expected number students in the Very Little category for each of the three streams, based on the assumption of independence.
- (3) There are more students in the 101 category overall, than would be expected under the assumption of independence.
- (4) There are more students in the Some category overall, than would be expected under the assumption of independence.
- (5) There are fewer 107 students in the Lots category, than would be expected under the assumption of independence.

1.65 versus  
3.17 & 4.25 & 6.92

restart on Ch 12 at  
2.40pm

**ANSWERS – CHAPTER 11**

- |         |         |         |         |         |         |
|---------|---------|---------|---------|---------|---------|
| 1. (4)  | 2. (4)  | 3. (5)  | 4. (4)  | 5. (4)  | 6. (2)  |
| 7. (2)  | 8. (3)  | 9. (3)  | 10. (1) | 11. (3) | 12. (1) |
| 13. (4) | 14. (3) | 15. (1) | 16. (4) | 17. (5) | 18. (3) |
| 19. (4) | 20. (5) | 21. (2) | 22. (5) | 23. (4) | 24. (5) |
| 25. (2) | 26. (1) |         |         |         |         |

711 Q5

# Revision Notes

straight lines

## Chapter 12 - Simple Linear Regression

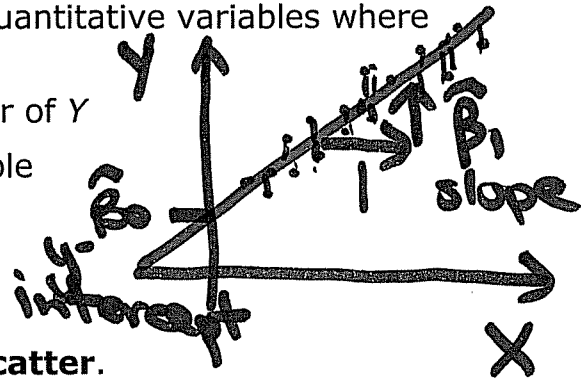
Look at blue pages for good notes and test/exam questions for practice

The main tool for comparing two quantitative variables is the scatter plot.  
What to look for in a scatter plot:

- o Trend (pattern)  $\rightarrow$  linear
- o Scatter  $\rightarrow$  non-linear
- o Outliers  $\rightarrow$  constant! (yes or no)
- o Association  $\rightarrow$  +ve  $\rightarrow$  or  $\rightarrow$  -ve
- o Strength of the relationship  $\rightarrow$  little  $\rightarrow$  strong, lots  $\rightarrow$  weak
- o Groupings

Regression looks at the relationship between two quantitative variables where the two variables take on special roles:

- X is used to **explain** or **predict** the behaviour of Y
- X is the **explanatory** or **independent** variable
- Y is the **dependent** or **response** variable



Two main components of the regression model are:

- o **trend** and
- o **scatter**.

We use a **least squares regression line**  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$  fitted by the computer / calculator to estimate the unknown population parameters  $\beta_0$  and  $\beta_1$

on formulae sheet

The single **least squares regression line** for each linear regression model:

- minimises the sum of the squared residuals/prediction errors
- has  $\sum \text{residuals} = 0$  (but so do many other lines)
- has  $(\bar{x}, \bar{y})$  lying on it

### Residuals

- Errors, residuals or prediction errors are all terms for the same thing.
- A residual is the (vertical) distance between the **actual observed value**  $y_i$  and the **expected estimated value**  $\hat{y}_i$ , i.e.:

Errors = observed - expected  $(\hat{u}_i = y_i - \hat{y}_i)$   
 or  
 memorise!

• **Hypotheses**

$H_0: \beta_1 = 0$  (there **is no linear relationship**)

$H_1: \beta_1 \neq 0$  (there **is a linear relationship**)



• **Assumptions** of simple linear regression are:

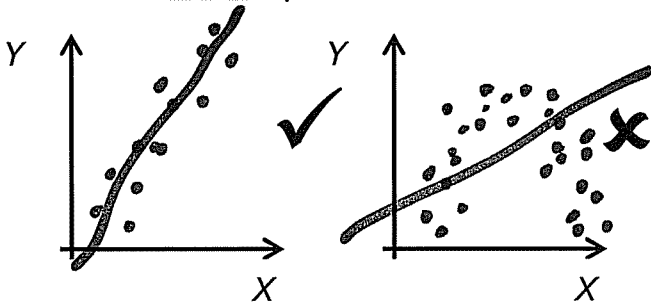
1. There is a **linear** relationship between  $X$  and  $Y$ .
2. Errors are **Normally** distributed (with  $\mu = 0$ ).
3. Errors all have the **same std deviation**,  $\sigma$ , regardless of the value of  $x$ .
4. Errors are all **independent**

*can't check this*

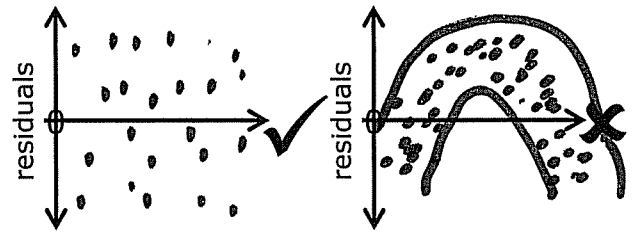
• **Assumption checking using plots of the data and residual plots**

1. There is a **linear** relationship between  $X$  and  $Y$ .

Scatterplot of data:

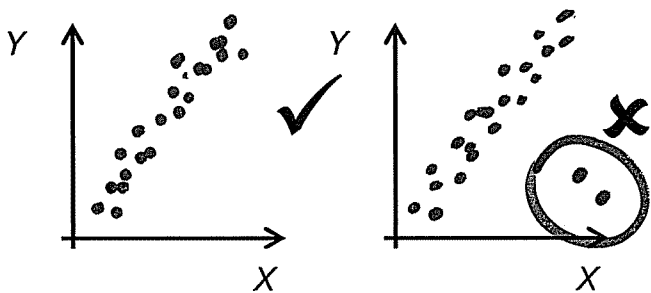


Residual plot:

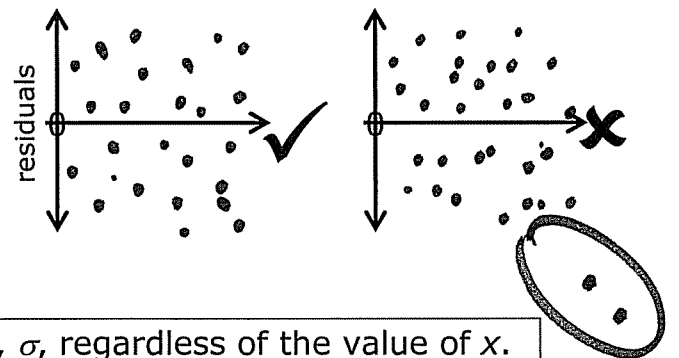


2. Errors are **Normally** distributed (with  $\mu = 0$ ).

Scatterplot of data:

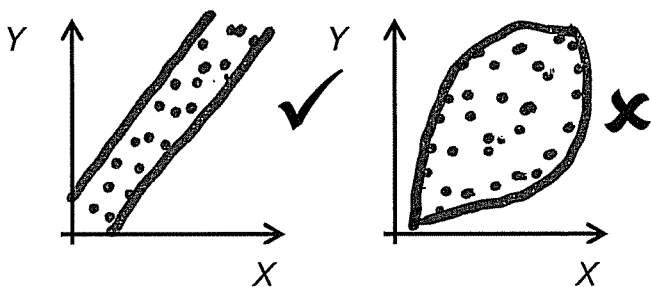


Residual plot:

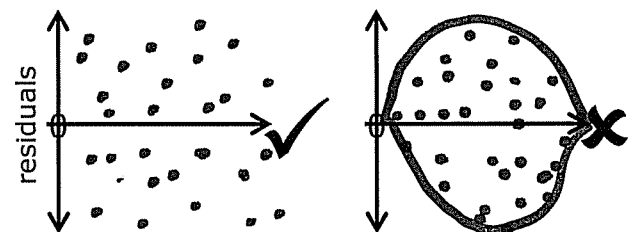


3. Errors all have the **same std deviation**,  $\sigma$ , regardless of the value of  $x$ .

Scatterplot of data:



Residual plot:



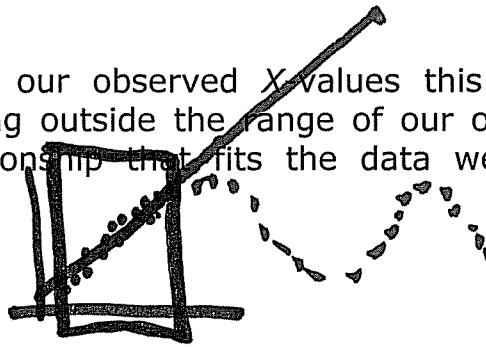
*4. errors are indep - can't check this!*

- **Degrees of Freedom**

$df = n - 2$  on formula sheet

- **Estimating / Predicting**

- Within the range of our observed  $X$ -values this can be done with confidence. Predicting outside the range of our observed  $X$ -values is dangerous. A relationship that fits the data well may not extend outside that range.



- ✓ **Confidence Interval (for the mean)**

This estimates the **mean**  $Y$ -value at a specified value of  $x$ .  
The width of the interval allows for:

- uncertainty about the values of  $\beta_0$  and  $\beta_1$ .

- ✓ **Prediction Interval**

This predicts the  $Y$ -value for **an individual** with a specified value of  $x$ .

The width of the interval allows for:

- uncertainty about the values of  $\beta_0$  and  $\beta_1$  **and**
- uncertainty due to the random scatter about the line.

- For a given value of  $x$ , the **95% prediction interval** is **always wider** than the **95% confidence interval for the mean**.

- **The Sample Correlation Coefficient,  $r$**

- $r$  measures the **strength** and **direction** of the **linear** association between **two quantitative variables**

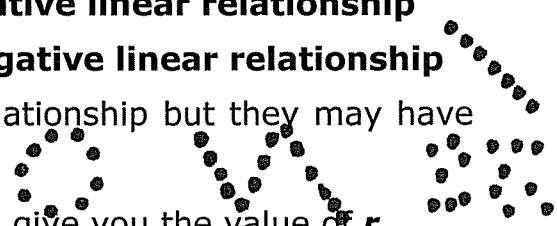
- The value of  $r$  is the same if the axes are swapped around – it doesn't matter which variable is  $X$  and which one is  $Y$  as  $r$  **treats both variables equally**

- $r$  measures how close the points in the scatter plot of  $Y$  against  $X$  (or vice versa) **come to lying on a straight line**

- $r = 1$ , then  $X$  and  $Y$  have a **perfect positive linear relationship**
- $r = -1$ , then  $X$  and  $Y$  have a **perfect negative linear relationship**
- $r = 0$ , then  $X$  and  $Y$  have **no linear** relationship but they may have **some other non-linear relationship**

- $r$  has no units – a computer / calculator can give you the value of  $r$

- **Correlation DOES NOT imply causation**



## Chapter 12 – Questions

1. The type of plot used to analyse variables in a regression model is a:

- (1) Side-by-side dot plot
- (2) Side-by-side box plot
- (3) Table of counts
- (4) Scatterplot
- (5) Histogram

2. Which one of the following statements is false?

- (1)  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are sample estimates of the parameters  $\beta_0$  and  $\beta_1$ .
- (2) The regression model consists of a linear trend plus random scatter.
- (3) A correlation coefficient, which is close to 0, indicates no linear relationship between  $X$  and  $Y$ .
- (4) The least squares estimates minimise the square of the difference between  ~~$\hat{x}$~~  and  ~~$\hat{y}$~~ .  $y_i$  &  $\hat{y}_i$
- (5) The errors from a regression analysis are assumed to be Normally distributed with a mean of zero.



3. The correlation between two variables,  $X$  and  $Y$ , is 0.7. Which one of the following statements is true?

- (1) There is ~~no~~ relationship between  $X$  and  $Y$ .
- (2) As the value of  $X$  increases, the value of  $Y$  tends to increase. *moderate to strong*
- (3) There is a ~~very weak~~ relationship between  $X$  and  $Y$ .
- (4) As the value of  $Y$  increases, the value of  $X$  tends to decrease.  $-0.7$
- (5) If the value of  $X$  is known then it is possible to determine the exact value of  $Y$ .

4. Which **one** of the following statements is **not** a reason for fitting a linear regression model to the data?

- (1) To estimate parameters in a theoretical model.
- (2) To make predictions.
- (3) To understand a relationship better.
- (4) To find the trend line.
- (5) To conclusively establish the cause of an effect.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Questions 5 to 8 refer to the following set of plots:

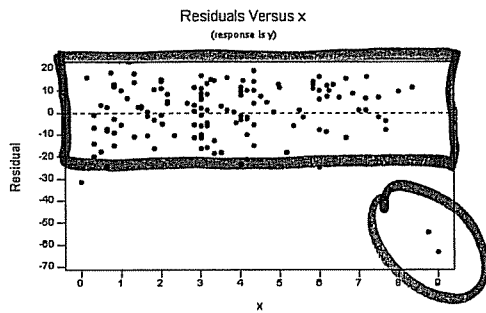


Figure A

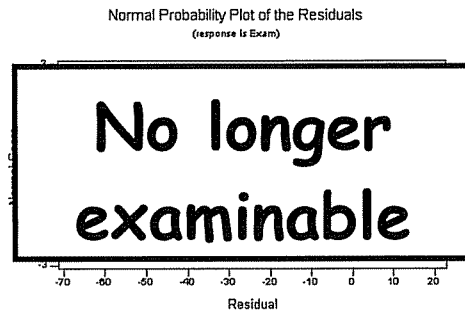


Figure B

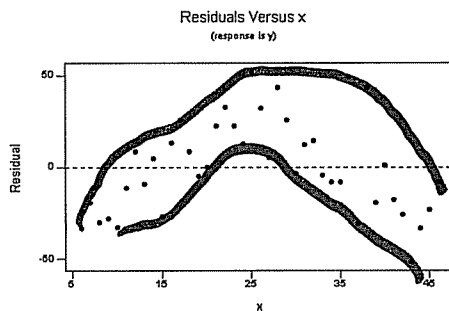


Figure C

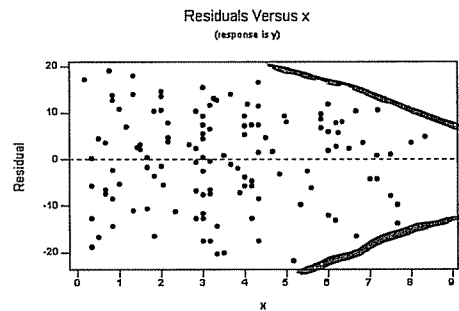


Figure D

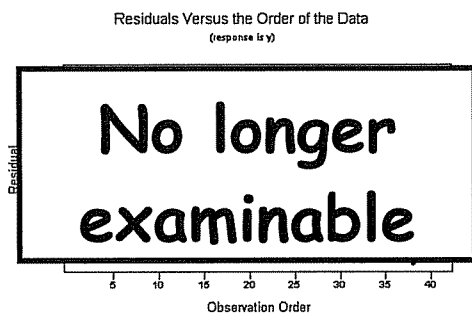


Figure E

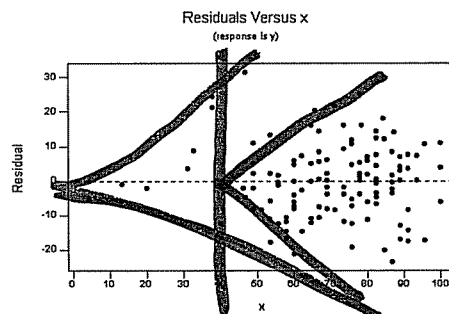


Figure F

In each of the above plots determine whether or not there any problems with the assumptions underlying linear regression model.

5. Figure A:

- (1) No problems. There is roughly a horizontal patternless band
- (2) Normality; Outliers
- (3) Non-linear
- (4) Non-constant scatter. This implies that the error variability is not independent of  $x$
- (5) Observations are not independent

6. Figure C:

- (1) No problems. There is roughly a horizontal patternless band
- (2) Normality; Outliers
- (3) Non-linear
- (4) Non-constant scatter. This implies that the error variability is not independent of  $x$
- (5) Observations are not independent

7. Figure D:

- (1) No problems. There is roughly a horizontal patternless band
- (2) Normality; Outliers
- (3) Non-linear
- (4) Non-constant scatter. This implies that the error variability is not independent of  $x$
- (5) Observations are not independent

8. Figure F:

- (1) No problems. There is roughly a horizontal patternless band
- (2) Normality; Outliers
- (3) Non-linear
- (4) Non-constant scatter. This implies that the error variability is not independent of  $x$
- (5) Observations are not independent

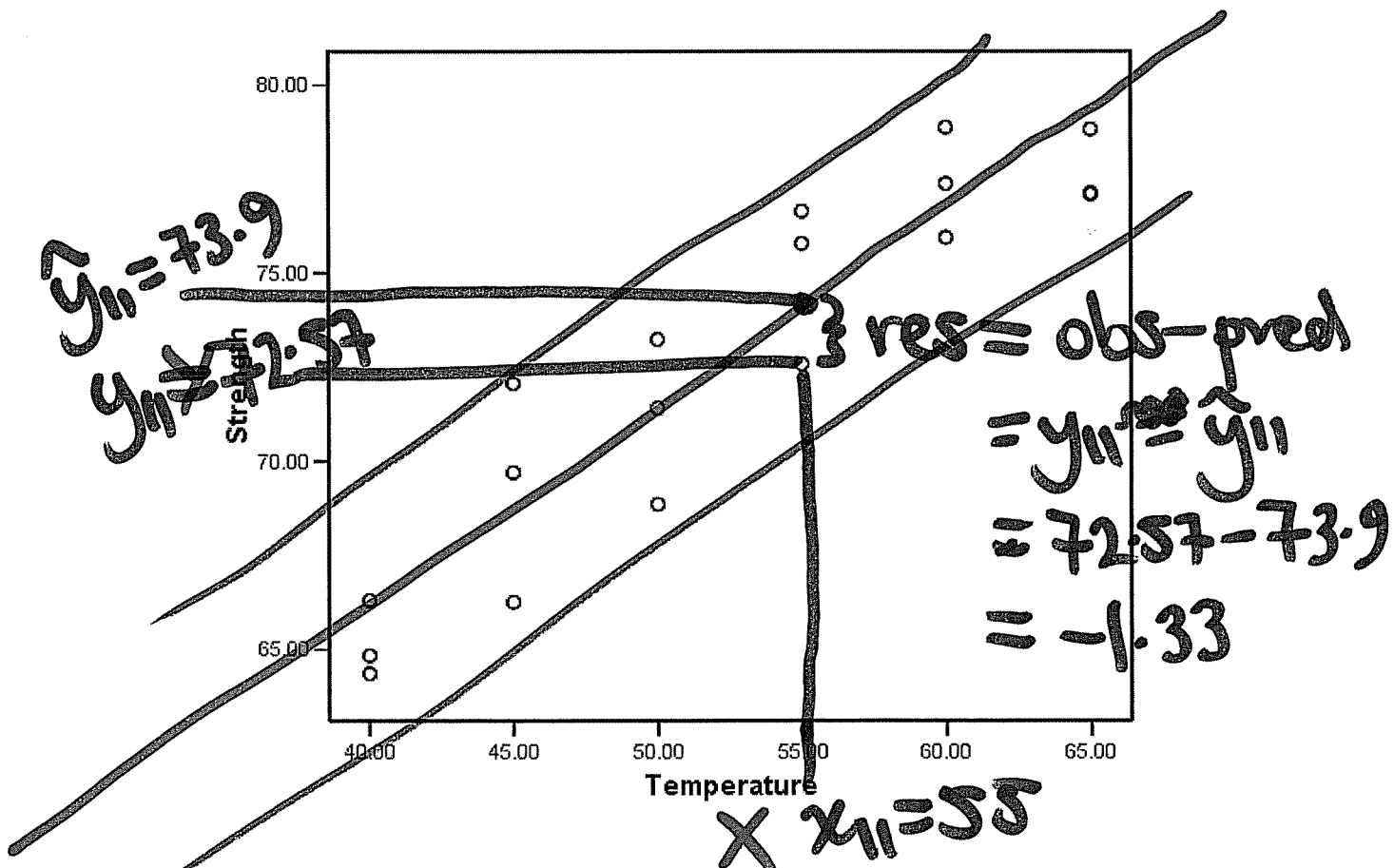
**Questions 20 to 28 refer to the following information.**

The production of particle boards involves a step in which the boards are baked. A manufacturer of particle boards investigated the effect of the baking temperature ( $X$ ) on the strength of particle boards ( $Y$ ). A total of 18 particle boards were baked using 6 different temperatures (3 boards were baked at each temperature) and the strength of these boards was measured. All aspects of the process other than the baking temperature were kept as similar as possible. The assignment of temperatures to boards and the order of production were determined using random processes. The data follows:

	Strength $Y$	Temperature ( $^{\circ}C$ ) $X$		Strength $Y$	Temperature ( $^{\circ}C$ ) $X$
1	66.30	40	10	75.78	55
2	64.84	40	11	72.57	55
3	64.36	40	12	76.64	55
4	69.70	45	13	78.87	60
5	66.26	45	14	77.37	60
6	72.06	45	15	75.94	60
7	73.23	50	16	78.82	65
8	71.40	50	17	77.13	65
9	68.85	50	18	77.09	65

A scatter plot and some computer output of these data are given below:

**Scatterplot of particle board strength against temperature**



Regression

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	45.454	2.846		15.972	.000
1 temperature	.518	.054	.924	9.672	.000

Dependent Variable: Strength

		Strength	Temperature
Strength	Pearson Correlation	1	.924(**)
	Sig. (2-tailed)		.000
	N	18	18
Temperature	Pearson Correlation	.924(**)	1
	Sig. (2-tailed)	.000	
	N	18	18

\*\* Correlation is significant at the 0.01 level (2-tailed).

y-intercept  
slope

$\hat{\beta}_0$   
 $se(\hat{\beta}_0)$   
 $\hat{\beta}_1$   
 $se(\hat{\beta}_1)$

$$t_0 = \frac{\hat{\beta}_0 - 0}{se(\hat{\beta}_0)} = \frac{45.454}{2.846}$$

$$H_0: \beta_0 = 0$$

$$H_0: \beta_1 = 0$$

$$t_0 = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{.518}{.054}$$

20. The fitted least squares regression line for these data is:

- ~~(1)  $\hat{y} = 45.454 + 0.518x$~~
- ~~(2)  $\hat{y} = 45.454 + 0.518$~~
- (3)  $\hat{y} = 45.454 + 0.518x$
- ~~(4)  $\hat{y} = 0.518 + 45.454x$~~
- ~~(5)  $\hat{y} = 0.518 + 45.454x$~~

pred. strength =  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$   
 $\hat{y} = 45.454 + .518x$   
 $\hat{y} = 45.454 + .518x \text{ Temp}$

21. The sample correlation coefficient for these data is:

- (1)  $r = 0.92$
- (2)  $r = 0.62$
- (3)  $r = 0.21$
- (4)  $r = 0.37$
- (5)  $r = 0.85$

22. The correct null and alternative hypotheses to test that there is no linear relationship between particle board strength and baking temperature are:

- ~~(1)  $H_0: \beta_0 = 0$  and  $H_1: \beta_0 \neq 0$~~
- (2)  $H_0: \beta_1 = 0$  and  $H_1: \beta_1 \neq 0$
- ~~(3)  $H_0: \beta_1 = 0$  and  $H_1: \beta_1 \neq 0$~~
- ~~(4)  $H_0: \beta_0 = 0$  and  $H_1: \beta_0 \neq 0$~~
- ~~(5)  $H_0: \beta_1 = 0$  and  $H_1: \beta_1 \neq 0$~~

versus  
or

23. To test the hypotheses from the previous question, the test statistic would be:

(1)  $t_0 = \frac{72.62}{2.846}$

(4)  $t_0 = \frac{0.518}{0.054}$

(2)  $t_0 = \frac{0.518}{2.846}$

(5)  $t_0 = \frac{45.454}{0.054}$

(3)  $t_0 = \frac{45.454}{2.846}$

$t_0 = \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)}$

24. When the hypothesis test referred to in questions 22 and 23 is conducted, it is found that there is very strong evidence against the hypothesis of no linear relationship between baking temperature and particle board strength. Which one of the following statements is true for this investigation?

$p\text{-val} = .000$

(1) ~~The results of this hypothesis test can be taken as evidence that changes in baking temperature cause changes in particle board strength since very strong evidence of a relationship always implies causation.~~

(2) ~~The results of this hypothesis test cannot be taken as evidence that changes in baking temperature cause changes in particle board strength since there may be factors other than baking temperature which affect the strength of particle boards.~~

(3) ~~The results of this hypothesis test cannot be taken as evidence that changes in baking temperature cause changes in particle board strength since strong evidence of a relationship does not necessarily mean the relationship is causal.~~

(4) ~~The results of this hypothesis test cannot be taken as evidence that changes in baking temperature cause changes in particle board strength since the scatter plot shows that for each baking temperature there is a substantial amount of variability in the strength of particle boards.~~

(5) The results of this hypothesis test can be taken as evidence that changes in baking temperature cause changes in particle board strength since a random process was used to assign boards to temperatures.

25. The fitted least squares regression line can be used to predict the strength of particle board. Boards baked at a temperature of 55°C have a predicted strength of approximately:

(1) 102.4

(4) 79.9

(2) 73.9

(5) 47.0

(3) 13.8

$\hat{y} = 45.451 + .518 \times 55$   
 $= 73.9$

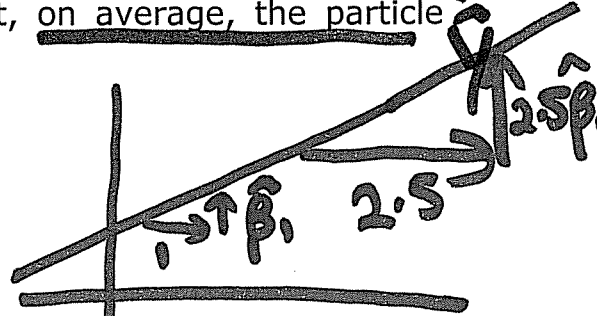
26. The baking temperature and particle board strength for sample number 11 was 55°C and 72.57 units respectively. Under the fitted least squares line, the value of the residual for this sample is approximately:

- (1) 2.30
- (2) 0.65
- (3) 1.33
- (4) -0.65
- (5) -1.33

$res = y_{11} - \hat{y}_{11} = 72.57 - 73.9 = -1.33$

27. The fitted least squares regression line indicates that for each increase of 2.5°C in baking temperature we expect that, on average, the particle board strength will:

- (1) increase by approximately 45.45 units.
- (2) decrease by approximately 0.52 units.
- (3) increase by approximately 0.52 units.
- (4) increase by approximately 1.30 units.
- (5) decrease by approximately 1.30 units.



$2.5 \times .518 = 1.295$

28. Why is the prediction interval for the strength of particle board baked at a temperature of 55°C more useful than the corresponding point estimate given in question 25?

- (1) Because the prediction interval is only one of the plausible values of that the strength could be when the baking temperature is 55°C.
- (2) The prediction interval is more useful than the point estimate as it estimates (with a certain level of confidence) a range of plausible values the strength could be when the baking temperature is 55°C.
- (3) Because the prediction interval is smaller than a corresponding confidence interval and therefore can capture the true value more accurately.
- (4) The prediction interval is more useful than the point estimate as it always captures the true estimate (with a certain level of confidence) in a range of plausible values the strength could be when the baking temperature is 55°C.
- (5) The prediction interval is not as useful as a corresponding confidence interval.

*larger*  
*value (pop. parameter)*

**ANSWERS – CHAPTER 12**

1. (4)	2. (4)	3. (2)	4. (5)	5. (2)	6. (3)
7. (1)	8. (4)	9. (3)	10. (2)	11. (2)	12. (4)
13. (2)	14. (3)	15. (5)	16. (5)	17. (3)	18. (5)
19. (5)	20. (3)	21. (1)	22. (2)	23. (4)	24. (5)
25. (2)	26. (5)	27. (4)	28. (2)		