Annual Conference 2007

Using IT in Statistics

Excel Spreadsheets

Bob Francis

- This session will look at the way Excel spreadsheets may be used to illustrate various statistical techniques in A Level Statistics modules.
- Each topic refers to an Excel spreadsheet which should be available to you.
- All of the 'data analysis' spreadsheets enable you to enter your own data for analysis.
- Alternatively you may choose a data set from the 'Data' sheet, which you can copy and paste on to the first page.
- You may find the spreadsheets useful for students of subjects such as Biology, Geography and Psychology, who just want a means of data processing to complete a task, project, etc.

Contents

Page	*Topic*
1	Introduction
2	Data Analysis — Discrete data
3	Data Analysis — Continuous data
4	Discrete Random Variables + Binomial Distribution
5	Approximating Distributions
6	Hypothesis Tester — Binomial and Poisson Distributions
7	Product Moment Correlation and hypothesis test
8	Spearman's Rank Correlation and hypothesis test
9	Least squares regression line and predictions
10	Chi-squared contingency table and test
11	Miscellaneous: Birthdays

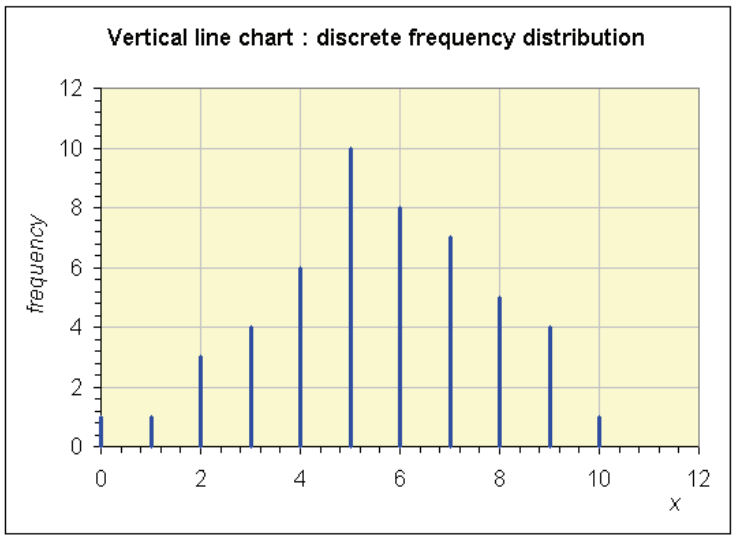
Data Analysis — Discrete Data

- You may enter a discrete data set with up to 12 pairs of data (x) and associated frequencies (f).
- Leave unused rows blank.
- Spreadsheet automatically calculates the values for the 'xf' and 'x^2f' columns, together with the column totals
- Spreadsheet displays the mean, standard deviation, root mean squared deviation and 'outlier' limits based on the 'two standard deviations from the mean' rule.
- Spreadsheet illustrates the data using a vertical line chart.

Data Analysis : Frequency Distribution : Vertical Line Chart

S1 book: Example 1.1, p19

	x	f	xf	x^2f
1	0	1	0	0
2	1	1	1	1
3	2	3	6	12
4	3	4	12	36
5	4	6	24	96
6	5	10	50	250
7	6	8	48	288
8	7	7	49	343
9	8	5	40	320
10	9	4	36	324
11	10	1	10	100
12				
	Totals	50	276	1770
	Mean =	5.52	s =	2.24
			$rmsd$ =	2.22



Mean - 2sd 1.03

Mean + 2sc 10.01

Over a period of time, a teacher recorded the number of times, x, each of the 20 students in the mathematics class was absent. The distribution was as follows.

Number of times absent, x	0	1	2	3	4	5	6	7	8	9	10	11 or more
Number of students, f	4	6	3	2	0	2	0	1	1	0	1	0

(i) Illustrate the data using a suitable diagram.

(ii) Calculate the mean and the standard deviation of the data set.

(iii) Determine if there are any outliers, and whether or not they should be excluded from the data set.

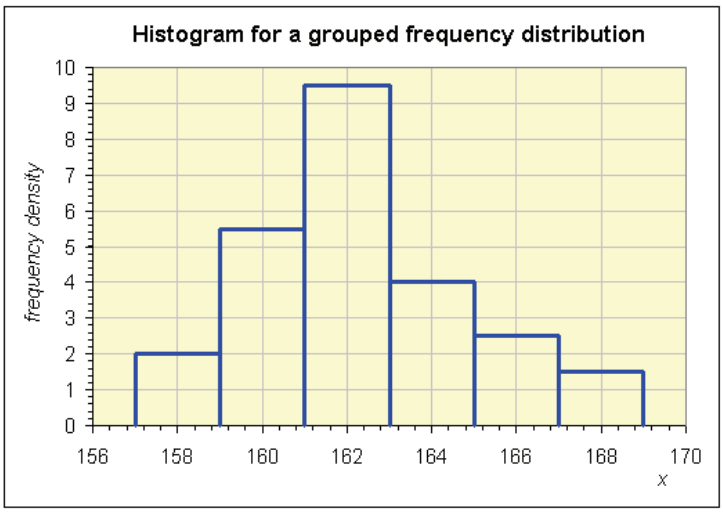
Data Analysis — Continuous Data

- You may enter a discrete data set with up to 12 pairs of data (*LCB and UCB*) and associated frequencies (*frequency*).
- Leave unused rows blank.
- Spreadsheet automatically calculates the values for the '*interval width*' and '*frequency density*' columns.
- Spreadsheet displays the mean, standard deviation, and 'outlier' limits based on the 'two standard deviations from the mean' rule. Calculations based on mid-interval '*x*' values.
- Spreadsheet illustrates the data using a histogram and a cumulative frequency curve (on separate page).

Data Analysis : Grouped Frequency Distribution : Histogram

S1 book, page 27 example

	LCB	UCB	frequency	int width	freq dens
1	157	159	4	2	2
2	159	161	11	2	5.5
3	161	163	19	2	9.5
4	163	165	8	2	4
5	165	167	5	2	2.5
6	167	169	3	2	1.5
7					
8					
9					
10					
11					
12					



Mean	162.32	s.d.	2.54	Mean - 2sd	157.24	Mean + 2sd	167.40

The amount spent, £x, by each of the 100 customers sampled is summarised in the following table.

Amount spent, £x	Number of customers
$0 \leq x < 10$	6
$10 \leq x < 20$	16
$20 \leq x < 30$	35
$30 \leq x < 50$	18
$50 \leq x < 75$	15
$75 \leq x < 100$	10

Mid-interval Frequency

x	f	fx	fx^2
158	4	632	99856
160	11	1760	281600
162	19	3078	498636
164	8	1312	215168
166	5	830	137780
168	3	504	84672
Totals	50	8116	1317712

(i) Illustrate the spending patterns by a histogram, drawn on graph paper.

(ii) Calculate estimates of the mean and standard deviation of the amount spent.

Discrete Random Variables

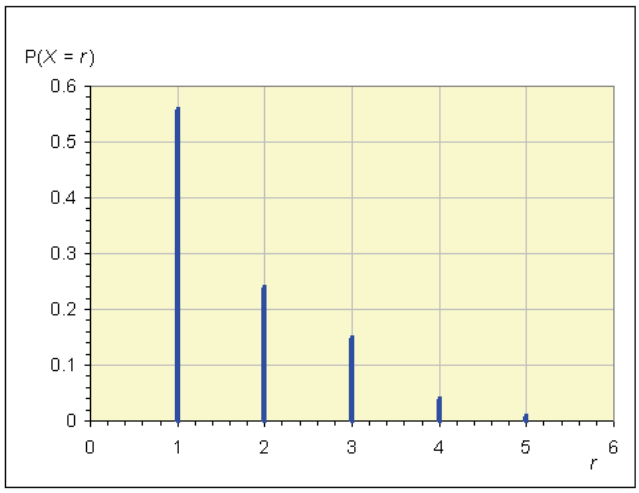
- You may enter a discrete data set with up to 12 pairs of data, r, and associated probabilities, P(X=r).
- Leave unused rows blank.
- Spreadsheet automatically calculates the values for the 'xf' and 'x^2f' columns, together with the column totals
- Spreadsheet displays the expectation E(X) and variance Var(X).
- Spreadsheet illustrates the data using a vertical line chart.
- Binomial Distribution sheet as a special case:
- Use slide bars to vary n and r.
- Examine relationship between n, r, E(X) and Var(X).

Discrete Random Variable : Expectation and variance : Vertical Line Chart

S1 Ch 4: Section 1

People per car (1)

	r	P(X= r)	r P(X= r)	r^2 P(X= r)
1	1	0.56	0.56	0.56
2	2	0.24	0.48	0.96
3	3	0.15	0.45	1.35
4	4	0.04	0.16	0.64
5	5	0.01	0.05	0.25
6				
7				
8				
9				
10				
11				
12				
	Totals	1	1.7	3.76
	E(X) =	1.700	Var (X) =	0.870



Expectation of X = E(X) = Σ r P(X = r)

Variance of X = Var(X) = E(X^2) - [E(X)]^2

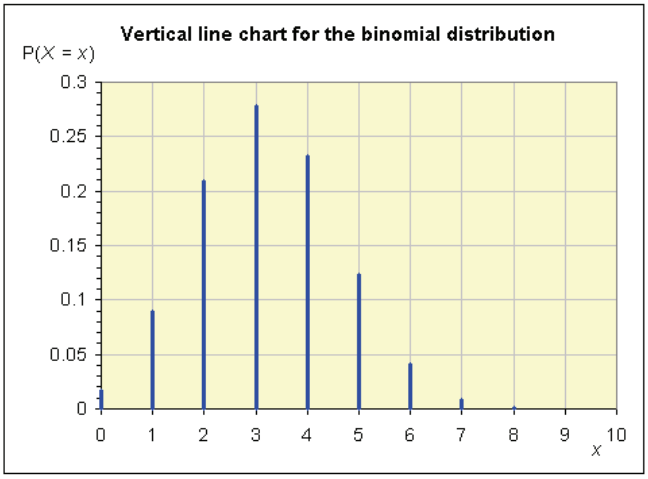
Discrete Random Variable : Expectation and variance : Vertical Line Chart

Binomial Distribution

n = 8 p = 0.4

r	P(X= r)	r P(X= r)	r^2 P(X= r)
0	0.01680	0.00000	0.00000
1	0.08958	0.08958	0.08958
2	0.20902	0.41804	0.83608
3	0.27869	0.83608	2.50823
4	0.23224	0.92897	3.71589
5	0.12386	0.61932	3.09658
6	0.04129	0.24773	1.48636
7	0.00786	0.05505	0.38535
8	0.00066	0.00524	0.04194
Totals	1	3.2	12.16

E(X) =	3.200	Var (X) =	1.920

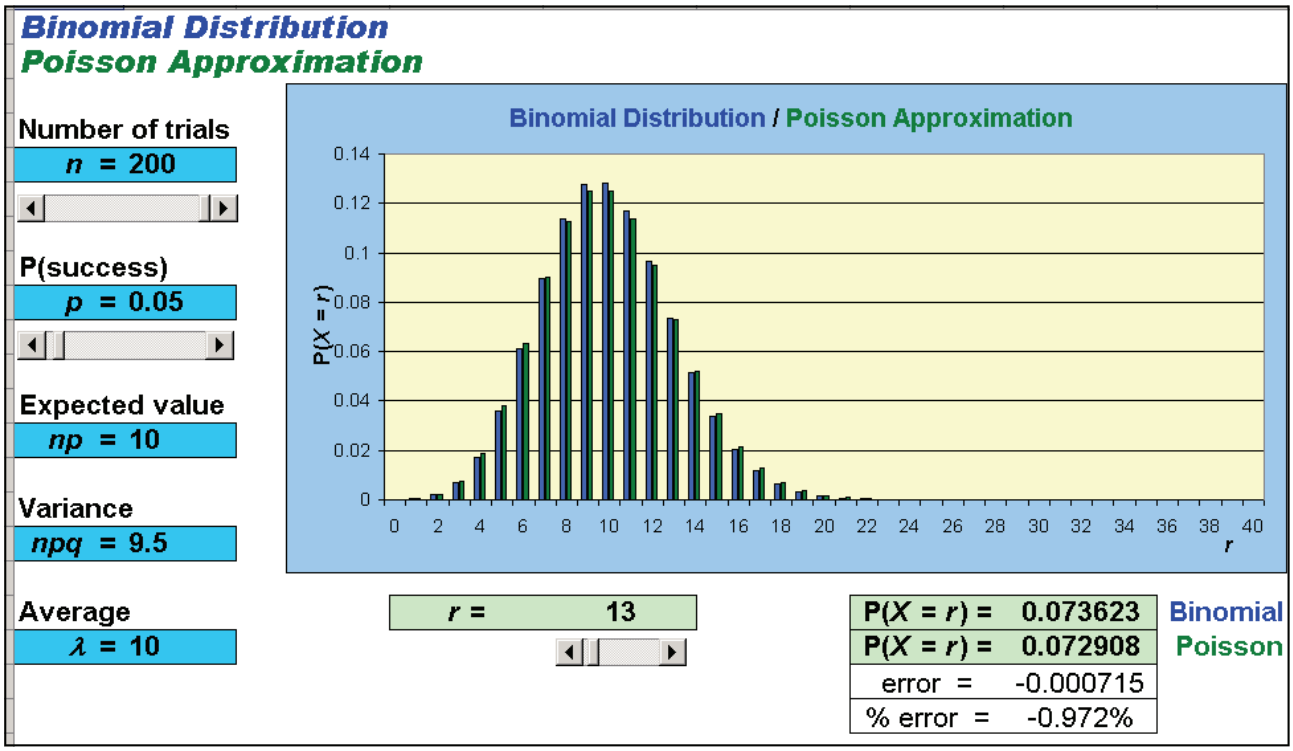


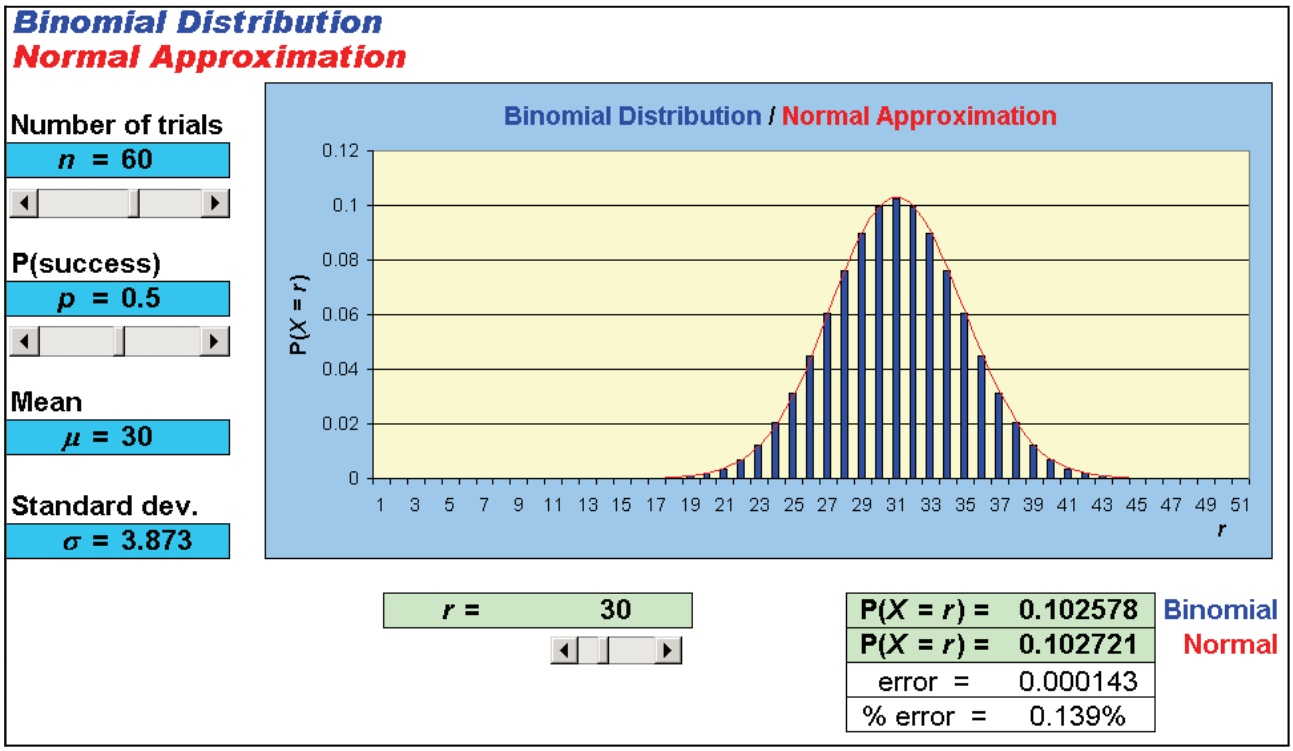
Expectation of X = E(X) = Σ x P(X = x)

Variance of X = Var(X) = E(X^2) - [E(X)]^2

Approximating Distributions

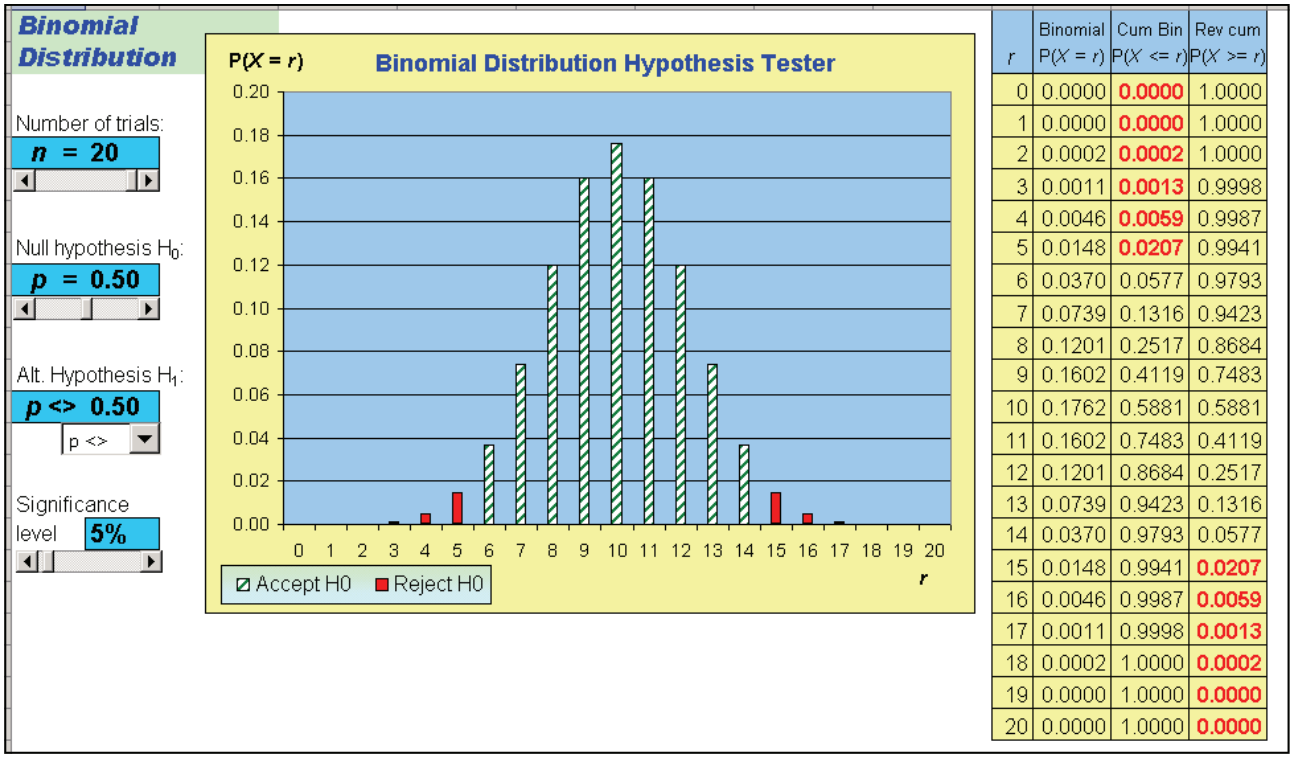
- Use slide bars to vary *n* and *p.*
- Poisson and Normal approximations to the Binomial Distribution
- Spreadsheet calculates Expectation, Variance and Average Rate (for Poisson)
- Spreadsheet calculates Mean and Standard Deviation (for Normal)
- Spreadsheet displays Binomial and approx. probabilities diagrammatically
- Use slide bar to vary *r*.
- Spreadsheet calculates P(*X* = *r*) exactly and approximately.





Binomial Distribution Hypothesis Tester

- Use slide bars to vary n and p (for Null hypothesis H_0).
- Use slide bar to vary form of Alternative Hypothesis H_1: one- or two-tail test.
- Use slide bar to vary significance level as a percentage.
- Spreadsheet calculates Binomial, Cumulative Binomial and Reverse Cumulative probabilities, colouring values of probabilities of 'critical region' r values in red.
- Spreadsheet displays Binomial probabilities diagrammatically, colouring bars of probabilities of 'critical region' r values in red.



r	Binomial P(X = r)	Cum Bin P(X <= r)	Rev cum P(X >= r)
0	0.0000	0.0000	1.0000
1	0.0000	0.0000	1.0000
2	0.0002	0.0002	1.0000
3	0.0011	0.0013	0.9998
4	0.0046	0.0059	0.9987
5	0.0148	0.0207	0.9941
6	0.0370	0.0577	0.9793
7	0.0739	0.1316	0.9423
8	0.1201	0.2517	0.8684
9	0.1602	0.4119	0.7483
10	0.1762	0.5881	0.5881
11	0.1602	0.7483	0.4119
12	0.1201	0.8684	0.2517
13	0.0739	0.9423	0.1316
14	0.0370	0.9793	0.0577
15	0.0148	0.9941	0.0207
16	0.0046	0.9987	0.0059
17	0.0011	0.9998	0.0013
18	0.0002	1.0000	0.0002
19	0.0000	1.0000	0.0000
20	0.0000	1.0000	0.0000

- A cross-channel ferry company runs a daily service from England to France.
- Records show that on average 85% of the ferry crossings leave on time, and the rest leave late.
- During the next three weeks there will be 21 departures.

(i) State a suitable distribution to model the number of times the ferry leaves on time and one assumption for the model to be valid.

(ii) For these three weeks, find the probability that

(A) all departures leave on time,

(B) exactly 3 departures leave late,

(C) during each week, no more than 1 of the 7 departures leaves late.

During the summer season it is suspected that fewer than 85% of ferry crossings will leave on time. In a random sample of 15 summer sailings, just 10 leave on time.

(iii) Carry out a suitable hypothesis test to examine the company's claim that 85% of summer sailings

Poisson Distribution Hypothesis Tester

The spreadsheet also contains a Poisson Distribution Hypothesis Tester, similar to the one for the Binomial Distribution. The test is for the parameter λ.

Pearson's PMCC and Hypothesis Tester

- You may enter a data set with up to 25 pairs of data (x, y).
- Use slide bar to vary form of Alternative Hypothesis H_1: one- or two-tail test.
- Use slide bar to vary significance level as a percentage.
- Spreadsheet calculates the sample correlation coefficient, r, and the critical value for the test, together with conclusion in terms of significance.
- Spreadsheet displays scatter diagram and a table of working.

PEARSON'S PRODUCT MOMENT CORRELATION COEFFICIENT

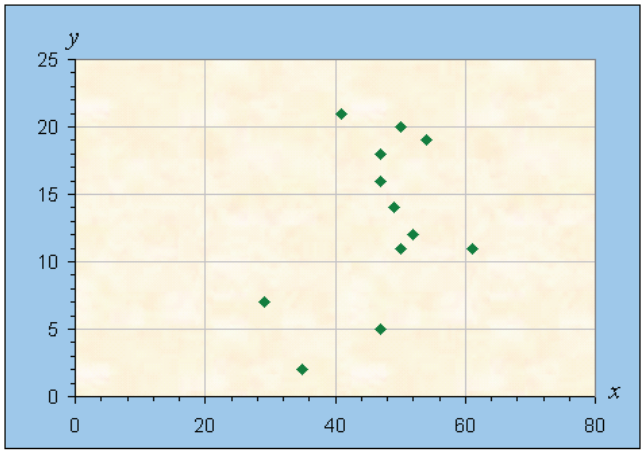
Place your data in cells B5 and C5 downwards; delete current data set and/or replace

Exercise 4B, q 2

	x	y
1	41	21
2	50	20
3	54	19
4	47	18
5	47	16
6	49	14
7	52	12
8	61	11
9	50	11
10	29	7
11	47	5
12	35	2

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{218}{573.9390212}$$

$r = 0.3798$



H_0: $\rho = 0$

H_1: $\rho > 0$ (>)

Significance level 10.0%

Critical value = 0.3981

Conclusion: Since 0.3798 < 0.3981 then result is **not significant** Accept H_0

	x	y	x^2	y^2	xy
1	41	21	1681	441	861
2	50	20	2500	400	1000
3	54	19	2916	361	1026
4	47	18	2209	324	846
5	47	16	2209	256	752
6	49	14	2401	196	686
7	52	12	2704	144	624
8	61	11	3721	121	671
9	50	11	2500	121	550
10	29	7	841	49	203
11	47	5	2209	25	235
12	35	2	1225	4	70

n =	12
Σx =	562
Σy =	156
Σx^2 =	27116
Σy^2 =	2442
Σxy =	7524

S_{xx} =	795.6667
S_{yy} =	414
S_{xy} =	218

A population analyst wishes to test how death rates and birth rates are correlated in European countries.

(i) State, with justification, appropriate null and alternative hypotheses for the test

A random sample of 10 countries from Europe was taken and their death rates (x) and birth rates (y), each per 1000 population for 1997, were noted.

x	9	9	7	12	11	10	7	13	8	7
y	14	9	13	13	10	11	16	9	16	12

(ii) Represent the data graphically.

(iii) Calculate the product moment correlation coefficient.

(iv) Carry out the hypothesis test at the 5% level of significance, stating any conclusion clearly .

Spearman's Rank CC and Hypothesis Tester

- You may enter a data set with up to 25 pairs of data (x, y).
- You may sort either the data (x, y) into *ascending* or *descending* order.
- Use slide bar to vary form of Alternative Hypothesis H_1: one- or two-tail test.
- Use slide bar to vary significance level as a percentage.
- Spreadsheet calculates the sample correlation coefficient, r_s, and the critical value for the test, together with conclusion in terms of significance.
- Spreadsheet displays scatter diagram and a table of working.

SPEARMAN'S RANK CORRELATION COEFFICIENT

Place your data in cells B5 and C5 downwards; delete current data set and/or replace

Exercise 4C, q 10

	x	y
1	1.54	1.7
2	1.5	3.95
3	1.49	2.75
4	1.22	1.97
5	1.19	2.35
6	1.11	1.45
7	1.09	2.4
8	1.06	2.05
9	1.05	2.15
10	0.97	2.3
11	0.88	1.75
12	0.68	2.1

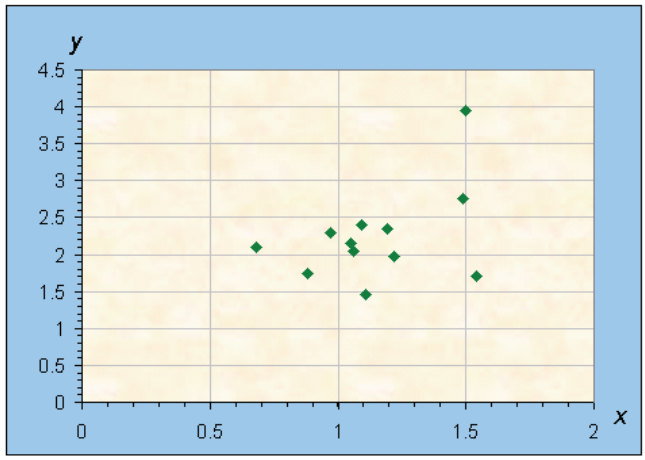
Sorting order: x Ascending; y Ascending

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{1440}{1716}$$

r_s = 0.1608

H_0: $\rho = 0$

H_1: $\rho > 0$



Significance level: 5.0%

Critical value = 0.5035

Conclusion: Since 0.1608 < 0.5035 then result is **not significant** — Accept H_0

	x	y	Rank x	Rank y	d	d^2
1	1.54	1.7	12	2	10	100
2	1.5	3.95	11	12	-1	1
3	1.49	2.75	10	11	-1	1
4	1.22	1.97	9	4	5	25
5	1.19	2.35	8	9	-1	1
6	1.11	1.45	7	1	6	36
7	1.09	2.4	6	10	-4	16
8	1.06	2.05	5	5	0	0
9	1.05	2.15	4	7	-3	9
10	0.97	2.3	3	8	-5	25
11	0.88	1.75	2	3	-1	1
12	0.68	2.1	1	6	-5	25

n = 12

Σd^2 = 240

$6\Sigma d^2$ = 1440

$n(n^2 - 1)$ = 1716

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} = 1 - \frac{1440}{1716}$$

Bird abundance may be assessed in several ways. In one long-term study in a nature reserve, two independent surveys (A and B) are carried out. The data show the number of wren territories recorded (survey A) and the numbers of adult wrens trapped in a fine mesh net (survey B) over a number of years.

Survey A	16	19	27	50	60	70	79	79	84	85	97
Survey B	11	12	15	18	22	35	35	71	46	53	52

(i) Plot a scatter diagram to compare results for the two surveys.

(ii) Calculate Spearman's coefficient of rank correlation.

(iii) Perform a significance test, at the 5% level, to determine whether there is any association between the results of the two surveys. Explain conclusion means in practical terms.

Least Squares Regression Line

- You may enter a data set with up to 25 pairs of data (x, y).
- Spreadsheet calculates $(\bar{x}, \bar{y})$ and the gradient of the regression line, b.
- Spreadsheet displays scatter diagram, with $(\bar{x}, \bar{y})$ highlighted.
- Spreadsheet displays a table of working, which includes predicted values, residuals and squares of residuals.
- Spreadsheet displays equation of regression line in two forms.

LEAST SQUARES REGRESSION LINE

Place your data in cells B5 and C5 downwards; delete current data set and/or replace

Exercise 4D, q 4

	x	y
1	0	0
2	1	3
3	2	6.8
4	3	10.2
5	4	12.9
6	5	16.4
7	6	20
8	7	21.4
9	8	23
10	9	24.6
11	10	26.1

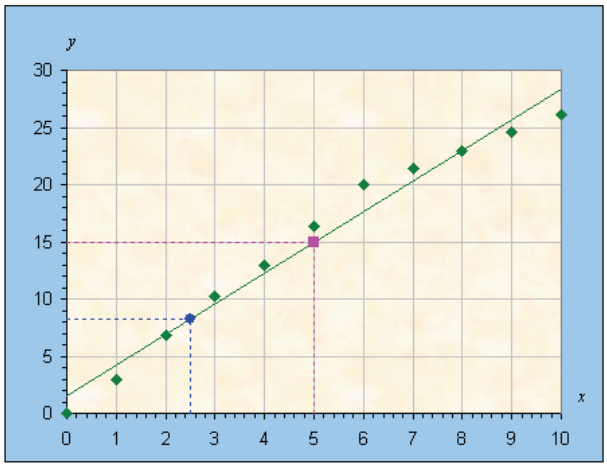
$\bar{x}$ = 5.00

$\bar{y}$ = 14.95

$b = \frac{S_{xy}}{S_{xx}} = \frac{295}{110} = 2.682$

x = 2.5

$\hat{y}$ = 8.24



Min x = 0 Max x = 10 Scale

Least squares regression line:

$y - 14.9 = 2.682(x - 5) \Rightarrow y = 2.682x + 1.5364$

	x	y	x^2	xy	$\hat{y}$	$\hat{y}-y$	$(\hat{y}-y)^2$
1	0	0	0	0	1.5364	1.53636	2.36041
2	1	3	1	3	4.2182	1.21818	1.48397
3	2	6.8	4	13.6	6.9	0.1	0.01
4	3	10.2	9	30.6	9.5818	-0.6182	0.38215
5	4	12.9	16	51.6	12.264	-0.6364	0.40496
6	5	16.4	25	82	14.945	-1.4545	2.1157
7	6	20	36	120	17.627	-2.3727	5.62983
8	7	21.4	49	149.8	20.309	-1.0909	1.19008
9	8	23	64	184	22.991	-0.0091	8.3E-05
10	9	24.6	81	221.4	25.673	1.07273	1.15074
11	10	26.1	100	261	28.355	2.25455	5.08298

n =	11
Σx =	55
Σy =	164.4
Σx^2 =	385
Σxy =	1117

$\bar{x}$ = 5

$\bar{y}$ = 14.945

S_{xx} =	110
S_{xy} =	295

An experiment was conducted to determine the mass, y g, of a chemical that would dissolve in 100 ml of water at x°C. The results of the experiment were as follows.

Temperature (x°C)	10	20	30	40	50
Mass (y g)	61	64	70	73	75

(i) Represent the data on graph paper.

(ii) Calculate the equation of the regression line of y on x. Draw this line on your graph.

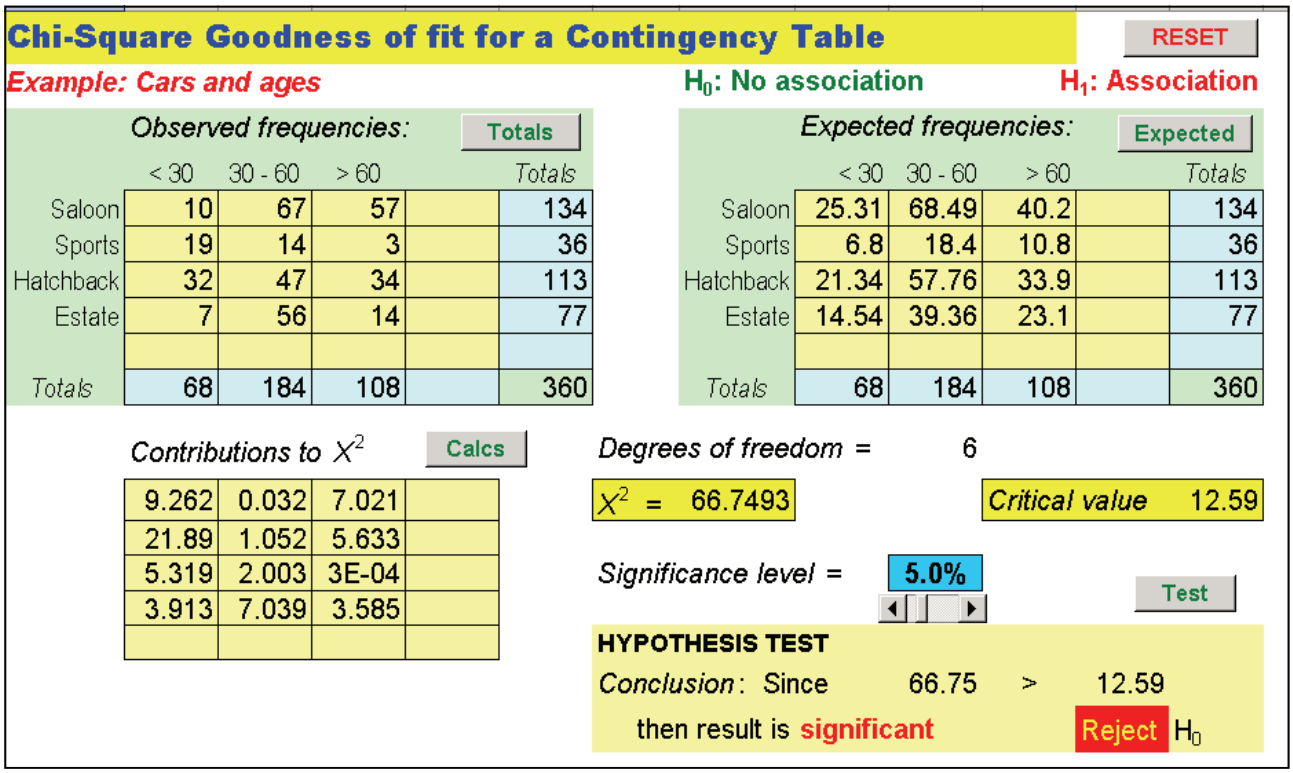
(iii) Calculate an estimate of the mass of the chemical that would dissolve in the water at 35°C.

(iv) Calculate the residuals for each of the temperatures. Illustrate them on your graph.

(v) Explain "least squares regression line" in relation to the residuals.

Chi-Square Contingency Table Tester

- You may enter observed frequencies into a table with up to 5 rows, 4 columns.
- Use slide bar to vary significance level as a percentage.
- Spreadsheet calculates the marginal totals for the observed frequencies, f_o.
- Spreadsheet calculates the table of associated expected frequencies, f_e.
- Spreadsheet calculates the table of contributions to the X^2 value.
- Spreadsheet calculates the X^2 value, the degrees of freedom and critical value.
- Spreadsheet gives the result of the significance test.
- You may step through the above process using the following buttons in order:
  RESET, Totals, Expected, Calcs, Test



The marketing manager at a theme park undertakes a survey of a random sample of 200 visitors. As part of the analysis, he categorises them as local people, people who have come a medium distance or people who have come a long distance, with a separate category of people in coach parties. He also categorises them according to the amount of money they spend in the park, as light, medium or heavy spenders.
A table displaying the results is as follows.

		Amount spent		
		Light	Medium	Heavy
Distance	Local	17	23	16
	Medium distance	15	25	34
	Long distance	4	16	12
	Coach party	8	22	8

(i) Stating your null and alternative hypotheses, examine whether or not there is any association between 'distance' and 'amount spent'. Use a 10% significance level.

(ii) Discuss your conclusions.

No. of people	P(all different birthdays)	P(at least two same)
1	1	0
2	0.997260	0.002740
3	0.991796	0.008204
4	0.983644	0.016356
5	0.972864	0.027136
6	0.959538	0.040462
7	0.943764	0.056236
8	0.925665	0.074335
9	0.905376	0.094624
10	0.883052	0.116948
11	0.858859	0.141141
12	0.832975	0.167025
13	0.805590	0.194410
14	0.776897	0.223103
15	0.747099	0.252901
16	0.716396	0.283604
17	0.684992	0.315008
18	0.653089	0.346911
19	0.620881	0.379119
20	0.588562	0.411438
21	0.556312	0.443688
22	0.524305	0.475695
23	0.492703	0.507297
24	0.461656	0.538344
25	0.431300	0.568700