Solutions of Cases of

Chapters 1 – 25

of the book

Statistical Methods for Business and Economics

by

Gert Nieuwenhuis
Tilburg University
The Netherlands

Solutions Cases Chapter 1

Solution Case 1.1

a. Exports: 3714.2×10^9 dollars; imports: 3791.0×10^9 dollars.

b. **Exports**:
Population 1 = 'World' = {Europe; North America; Asia; Middle East; Africa; CIS; South and Central America}.
Population 2 = 'Economies' = {European Union (25); United States; Switzerland; China; Russian Federation; Japan; Turkey; Norway; Canada; Australia; China Hong Kong; United Arab Emirates; Romania; Republic of Korea; India; Singapore; South Africa; Mexico; Brazil; Chinese Taipei; Israel; Saudi Arabia; Islamic Rep. of Iran; Ukraine; Croatia; Algeria; Morocco; Malaysia; Tunisia; Egypt}.
Imports:
Population 1 = 'World' = {Europe; Asia; North America; CIS; Africa; South and Central America; Middle East}.
Population 2 = 'Economies' = {European Union (25); United States; China; Japan; Russian Federation; Switzerland; Norway; Turkey; Republic of Korea; Chinese Taipei; Brazil; Singapore; India; Canada; Saudi Arabia; Malaysia; South Africa; Romania; Libyan Arab Jamahiriya; Thailand; Algeria; Indonesia; China Hong Kong; Australia; Israel; Islamic Rep. Of Iran; Chile; Ukraine; Mexico; Tunisia}.

c. They do not cover the whole world, not for exports and also not for imports: the population-totals of the trading activity is not equal to the exports total (respectively the imports total) of part **a**.

d. **Exports** (to an economy):
   1. amount of exports in 2004 (billion dollars) to the economy
   2. percentage exported to the economy in 2000
   3. percentage exported to the economy in 2004
   4. change in exports in 2003 to the economy when compared to the year before (as a percentage)
   5. change in exports in 2004 to the economy when compared to the year before (as a percentage)

   Imports (from an economy):
   1. amount of imports in 2004 (billion dollars) from the economy
   2. percentage imported from the economy in 2000
   3. percentage imported from the economy in 2004
   4. change in imports from the economy in 2003 when compared to the year before (measured as a percentage)
   5. change in imports from the economy in 2004 when compared to the year before (measured as a percentage)

e. **Exports**:
Top 5: the first five of the exporting economies in population 2;
the rest: the last 25 exporting economies of population 2.

Variable	**1**	**2**	**3**	**4**	**5**
Top 5	3007.6	80.9	81.0	20	18
Rest	511.9	13.8	13.8	--	--

Imports:

Top 5: the first five of the importing economies in population 2;
the rest: the last 25 importing economies of population 2.

Variable	**1**	**2**	**3**	**4**	**5**
Top 5	3043.0	79.3	80.3	20	19
Rest	547.0	14.8	14.4	--	--

Solutions Cases Chapter 2

Solution Case 2.1 See book

Solution Case 2.2

a. - In 2001: 1537; 761 males and 776 females

-

	2001		
Age group	**Males**	**Females**	**Total**
0-4	112	106	218
5-9	118	88	206
10-14	102	100	202
15-19	71	73	144
20-24	42	43	85
25-29	43	52	95
30-34	50	50	100
35-39	43	45	88
40-44	41	43	84
45-49	28	35	63
50-54	20	32	52
55-59	26	28	54
60-64	21	28	49
65+	44	53	97
Total	761	776	1,537

-

Population 5 years and over by highest qualification gained at school and sex, Tokelau, 2001.

Highest qulaifications gained at school	Male	Female	Total
None	268	253	521
Primary/Form 2 Certificate	73	79	152
Leaving Certificate	43	54	97
School Certifcate	45	47	92
University Entrance	13	21	34
Other	5	4	9
Total	447	458	905

b. - Currency: New Zealand dollars

-

Year	**Imports**
1999	1,110,152
2000	1,762,310
2001	1,846,083
2002	2,087,696
2003	373,932
2004	174,190

Total imports value, major items	**2002**
Total	1,673,389
Food & Live Animal	923,766
Beverages & Tobacco	275,915
Mineral fuels, Lubricants & Related Materials	194,779
Animal & Vegetable Oils, Fats & Waxes	50,012
Chemicals & Related Products	45,429
Manufactured Goods Classified Chiefly by Material	55,273
Miscellaneous Manufactured Goods & Articles	128,215

- During the period 1999 – 2004 the export was 0.

c. - In 2001:

Kind of work done a week before the census	Male	Female	Total
Fishing, gardenning, handicraft, bread-baking or making Todday	48	49	97
Only other type of work	274	167	441
A combination of the above	4	0	4
No work	80	244	324
Total	406	460	866

Not very informative. The distributions of the males and the females differ; for instance: much more women have no work.

- in 2001:

Occupation (major Groups)	Male	Female	Total
Religious	4	0	4
Labourers?cleaners	101	64	165
Carpenters/Builders	79	2	81
Computer IT specialists	4	1	5
Electrical/Other technicians	13	3	16
Medical professionals	4	16	20
Sevice workers	16	15	31
Administrative/Clerical workers	36	32	68
Teachers	15	29	44
Politicians	6	0	6
Total	278	162	440

The distributions of the males and the females differ. For instance, men relatively often are labourers/cleaners; females relatively often are teachers.

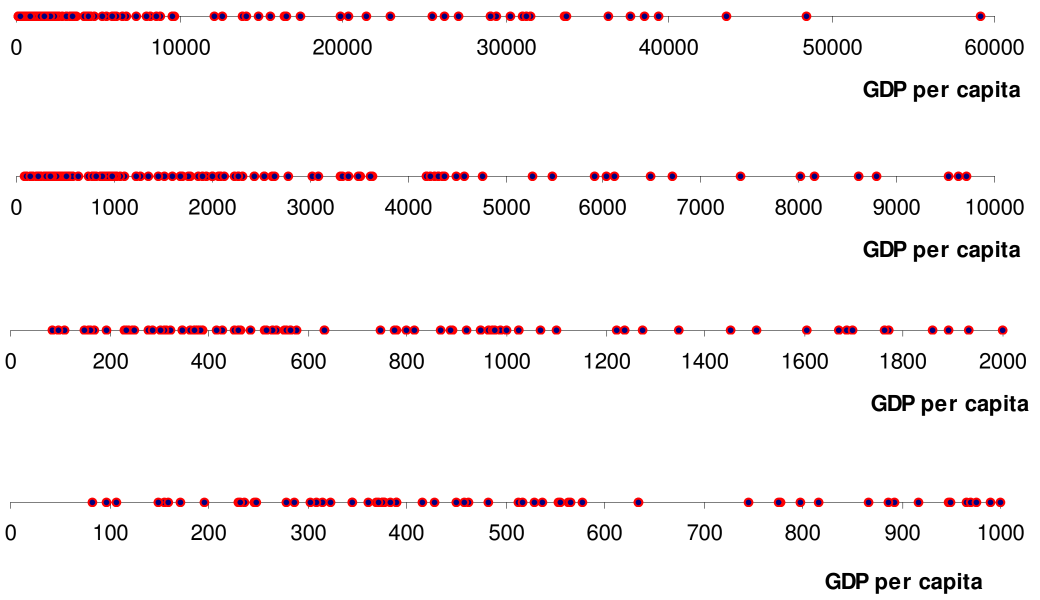
- In 2001:

Industry (Major Groups)	Male	Female	Total
Construction	76	2	78
Retail Trade	8	4	12
Hotels, Restaurants	2	2	4
Transport	6	1	7
Communication/Other services	10	10	20
Village Services	115	67	182
Public Administration	33	26	59
Education	21	32	53
Medical, Dental	5	18	23
	276	162	438

Solution Case 2.3

a.

Figure. Dot plots of GDP per capita (2003) for 168 countries.



Many countries are – as far as their GDPpc is concerned – located between 0 and 1000. After GDPpc-level 1000, the density of the dots dies out and is very low after 30000.

b.

Table. Classified GDP per capita.

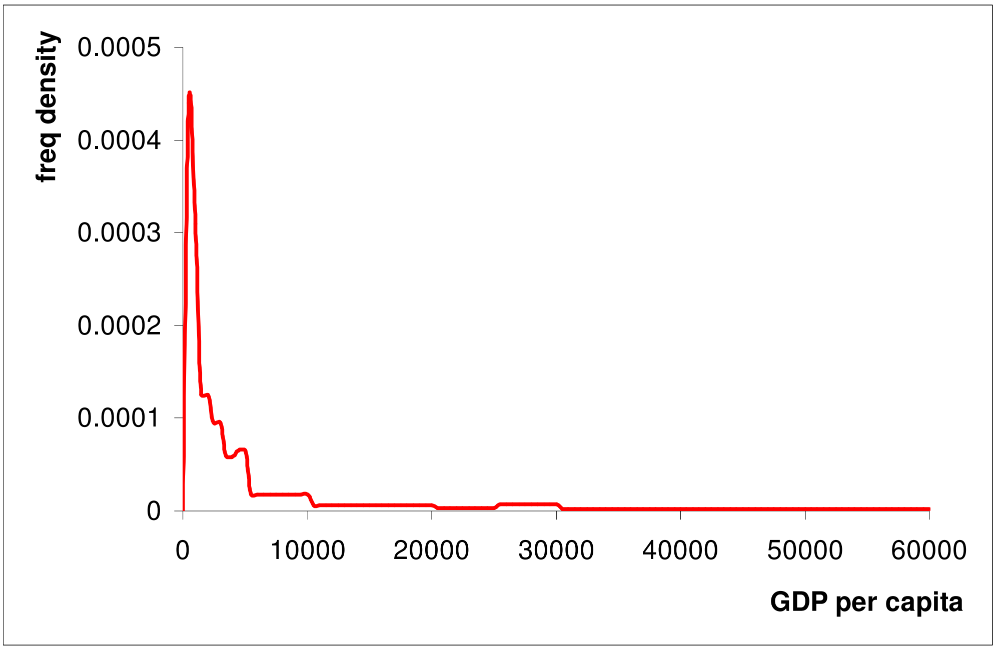
Class	Frequency	Rel frequency	Frequency density
(0, 500]	37	0.22024	0.000448
(500, 1000]	26	0.15476	0.000310
(1000, 2000]	21	0.12500	0.000125
(2000, 3000]	16	0.09524	0.000095
(3000, 4000]	10	0.05952	0.000060
(4000, 5000]	11	0.06548	0.000065
(5000, 10000]	15	0.08929	0.000018
(10000, 15000]	5	0.02976	0.000006
(15000, 20000]	5	0.02976	0.000006
(20000, 25000]	3	0.01786	0.000004
(25000, 30000]	6	0.03571	0.000007
(30000, 60000]	13	0.07738	0.000003
Total	168	1	-----

Original source: United Nations Development Report (2006)

Notice that the classes (0, 500] and (500, 1000] jointly include the GDPpc of 37.5% of the countries, while their joint width 1000 is only 1.7% of the width of the total range 0 – 60000. Furthermore, the frequency density of each of the classes from GDPpc-level 10000 onwards is at most 1.6% of the frequency density of the class (0, 500]. These facts illustrate how inequitably wealth is distributed among the countries in the world.

c. The graph in the picture below is called a ***scatter plot***. It shows how the frequency density (on the vertical axis) is related to GDP per capita on the horizontal axis.

Figure. Scatter plot of the frequency density on GDP per capita.



This plot is chosen instead of a histogram since histogram-bars that belong to the narrow, lower classes would disturb the picture.

Solutions Cases Chapter 3

Solution Case 3.1 See book

Solution Case 3.2

a. The integers a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2 are respectively equal to 1, 2, 5, 8, 6, 4, 8, 5. Indeed: 1/5 < 2/8 and 6/8 < 4/5, while (1+6)/(5+8) > (2+4)/(8+5).
b. No, since all ratios remain unchanged.
c. Variable: X = 'number of goals per match'. There are six means involved:

- Makaay, on the population of all matches he plays or played for the Dutch team: the mean of the five sample observations is 1/5.
- Makaay, on the population of all matches he plays or played for his private employer: the mean of the eight sample observations is 6/8.
- Van Nistelrooij, on the population of all matches he plays or played for the Dutch team: the mean of the eight sample observations is 2/8.
- Van Nistelrooij, on the population of all matches he plays or played for his private employer: the mean of the five sample observations is 4/5.
- Makaay, on the population of all matches he plays or played for the Dutch team or for his private employer: the mean of the 13 sample observations is 7/13.
- Van Nistelrooij, on the population of all matches he plays or played for the Dutch team or for his private employer: the mean of the 13 sample observations is 6/13.

d. The integers a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2 are respectively equal to 70, 15, 100, 20, 5, 35, 20, 100. Indeed: 70/100 < 15/20 and 5/20 < 35/100, while (70+5)/(100+20) > (15+35)/(20+100).
e. Variable X is a 0 - 1 variable: it takes the value 1 if a candidate is invited and the value 0 if not. So, the six means are sample fractions. The populations are the six sets of male (or female) candidates on Harvard and/or Oxford.

Solution Case 3.3

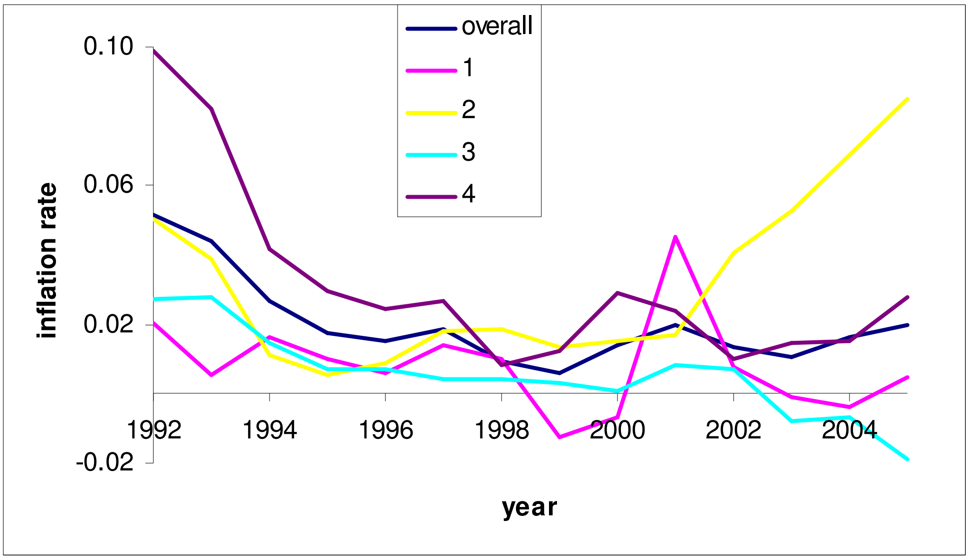
a.

Table. Inflation rates Germany, 2001 – 2005.

Year	2001	2002	2003	2004	2005
Inflation rate (%)	0.0200	0.0137	0.0106	0.0163	0.0198

b. mean = 0.0161;
geometric mean = $\sqrt[5]{1.0200 \times 1.0137 \times \cdots \times 1.0198} - 1 = 0.0161$.
c. Use the possibilities of your statistical package to create the inflation data; see Appendix A1.
d. Arithmetic means: 0.02023; 0.00823; 0.03168; 0.00553; 0.03174
Geometric means: 0.02016; 0.008141; 0.031407; 0.00546; 0.03143
e.
2 = Alcoholic beverages, tobacco;
4 = Housing, water, electricity, gas and other fuels
f. $81.9 \times (1+0.02016)^{14} = 108.3$

g.



h. From 2002 onwards, the prices in the sector alcoholic beverages and tobacco are certainly rising. On the other hand, the prices in the sector food, etc. go down. The overall picture doesn't seem to be influenced by the introduction of the euro.

Solutions Cases Chapter 4

Solution Case 4.1 See book

Solution Case 4.2

a. The table below is part of the Descriptive Statistics printout of Excel:

Mean	21.8750
Median	19.5000
Standard deviation	16.4257
Sample variance	269.8045
Range	63.0000
Minimum	0.0000
Maximum	63.0000
Sum	875.0000
Count	40.0000

Note that the standard deviation and the variance are sample statistics. To get the population variance, multiply 269.8045 by 39/40 to obtain 263.0594. Taking the square-root gives 16.2191, the population standard deviation.

b. See **a**.

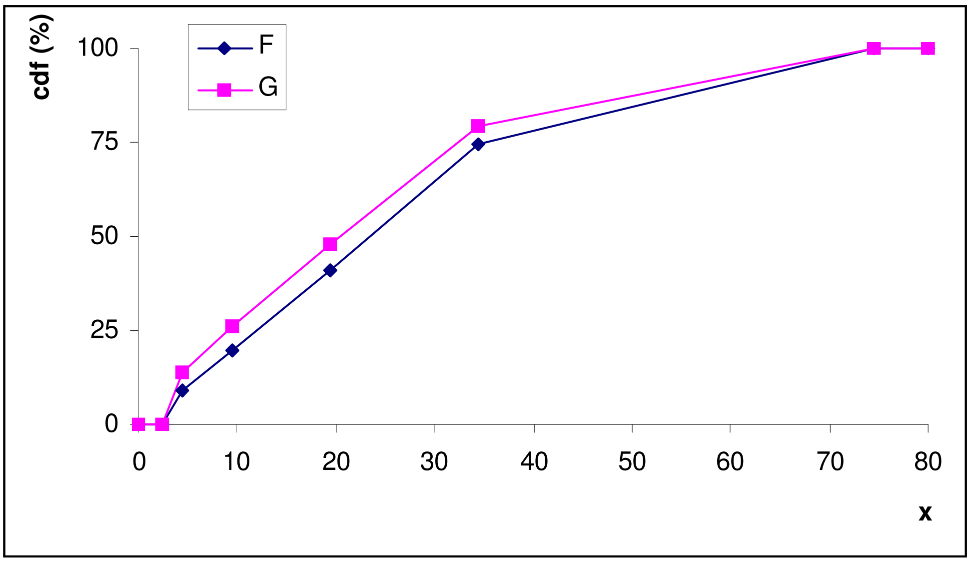
c. Queen Victoria reigned 63 years, the maximum. Edward V reigned 0 years (rounded). It means that this king reigned less than 0.5 years.

d. 3σ-interval: (–26.7823, 70.5323). Hence, Elizabeth II has to reign (as seen from 2007 onwards) another 15-16 years to become an outlier in this sense.
Excel: $\kappa_1 = 9.75$ and $\kappa_3 = 33.5$
1.5δ-interval: (–25.875, 69.125). Hence, Elizabeth II has to reign (as seen from 2007 onwards) another 14-15 years to become an outlier in this sense.

Solution Case 4.3

a. Note that X is a continuous variable. The fourth columns of the two (adjacent) tables give the levels of the cdf's at the end of the classes. Here are the graphs:

Figure. Graphs of the cdf's F and G.



It follows that $F(x) \le G(x)$ for all *x*. This illustrates that, at least at first view, there are some positive developments: the 2000/2002-situation is more concentrated on lower values *x* of the 'hunger'-variable *X*.

b. The table recalls the two frequency distributions:

		1990-1992		2000-2002	
Class	**Centre**	**Frequency**	***F* (in %) at endpoint**	**Frequency**	***G* (in %) at endpoint**
2.5 - 4.5	3.5	8	9.3	12	13.6
4.5 - 9.5	7	9	19.8	11	26.1
9.5 - 19.5	14.5	18	40.7	19	47.7
19.5 - 34.5	27	29	74.4	28	79.5
34.5 - 74.5	54.5	22	100	18	100
Total	---	86	---	88	---

The (approximating) mean values follow by multiplying frequencies and centres, adding up the results and dividing the sum by the corresponding size of the population:

$$\text{1990/1992:} \quad \frac{8\times 3.5+\cdots+22\times 54.5}{86} = \frac{2334}{86} = 27.14$$

$$\text{2000/2002:} \quad \frac{12\times 3.5+\cdots+18\times 54.5}{88} = \frac{2131.5}{88} = 24.22$$

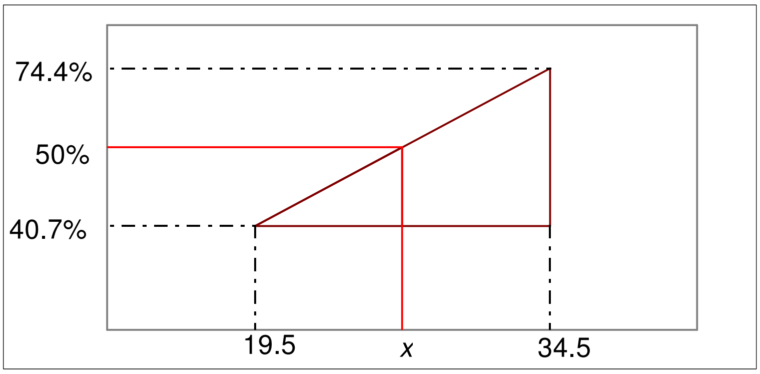
Hence, the mean percentage (per developing country) of undernourished inhabitants decreased slightly from 27% to 24%.

To determine the medians, the equations $F(x) = 50\%$ and $G(x) = 50\%$ have to be solved. Notice that both *F* and *G* pass the vertical level 50% between 19.5 and 34.5:

$$F(19.5) = 40.7 \quad \text{and} \quad F(34.5) = 74.4\,;$$
$$G(19.5) = 47.7 \quad \text{and} \quad G(34.5) = 79.5$$

Since *X* is continuous, *F* and *G* increase linearly between 19.5 and 34.5. The graph below is a rough sketch that is used to solve $F(x) = 50\%$; note that the proportion along the vertical axis are denoted as percentages.

Calculation of solution *x* of *F*(*x*) = 0.5 (50%)



By linear interpolation with the help of the triangle-construction it follows that:

$$\frac{x-19.5}{34.5-19.5}=\frac{50-40.7}{74.4-40.7}\text{, which yields } x = 23.6;$$

$$\frac{x-19.5}{34.5-19.5}=\frac{50-47.7}{79.5-47.7}\text{, which yields } x = 20.6$$

Hence, the medians of F and G are respectively 23.6% and 20.6%.

To determine the modal class, the classes with maximal frequency **densities** have to be obtained. For both populations, this is the class (2.5, 4.5]. Hence – for both cdf's – mean, median and mode are ordered as follows:

mode < median < mean

Apparently large observations force the mean to be larger than the median, also in 2000/2002. Note that mean and median both decreased 3 units between 1990/1992 and 2000/2002. On the other hand, the number of developing countries with more than 2.5% undernourished increased from 86 to 88.

c. The two population variances can be calculated with the short-cut formula:

1990/1992: $\dfrac{8\times 3.5^2+\cdots+22\times 54.5^2}{86}-27.14^2 = 319.3506$

2000/2002: $\dfrac{12\times 3.5^2+\cdots+18\times 54.5^2}{88}-24.22^2 = 306.0876$

Hence, the standard deviation decreased only slightly from 17.87 to 17.50

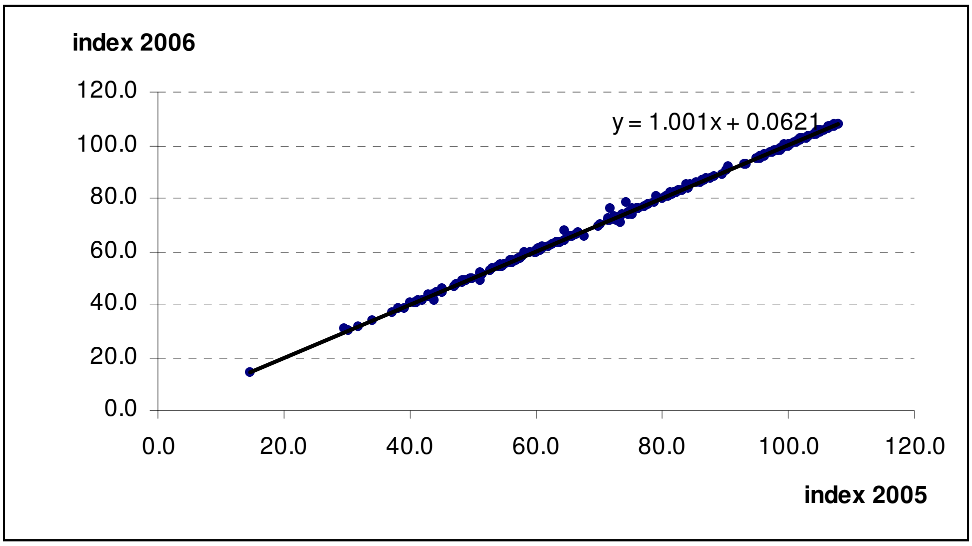
d. The facts that in general G is concentrated at smaller values than F and that both the mean and the standard deviation have decreased, are positive developments.

Solutions Cases Chapter 5

Solution Case 5.1 See book

Solution Case 5.2

a.



b. 0.999571; there is a very strong positive linear relation.

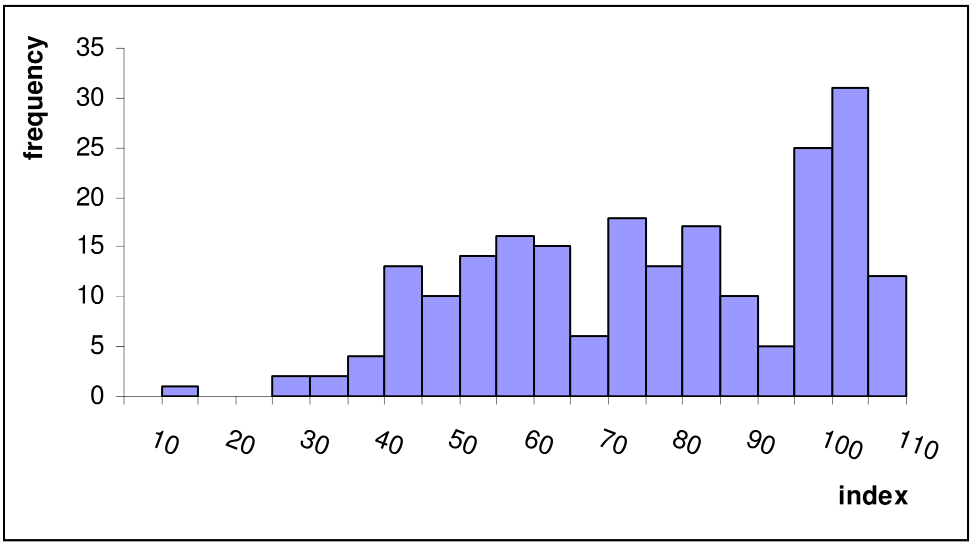
c. Kuwait, Tel Aviv, Jerusalem

d. $\hat{y} = 1.001 \times 105.7 + 0.0621 = 105.868$

e.

Bin	15	20	25	30	35	40	45	50	55	60
Freq	1	0	0	2	2	4	13	10	14	16
Perc	0.467	0	0	0.935	0.935	1.869	6.075	4.673	6.542	7.477
Cum perc	0.47	0.47	0.47	1.405	2.339	4.208	10.28	14.96	21.5	28.97
Bin	65	70	75	80	85	90	95	100	105	110
Freq	15	6	18	13	17	10	5	25	31	12
Perc	7.009	2.804	8.411	6.075	7.944	4.673	2.336	11.68	14.49	5.607
Cum perc	35.98	38.79	47.2	53.27	61.22	65.89	68.23	79.91	94.4	100

f.

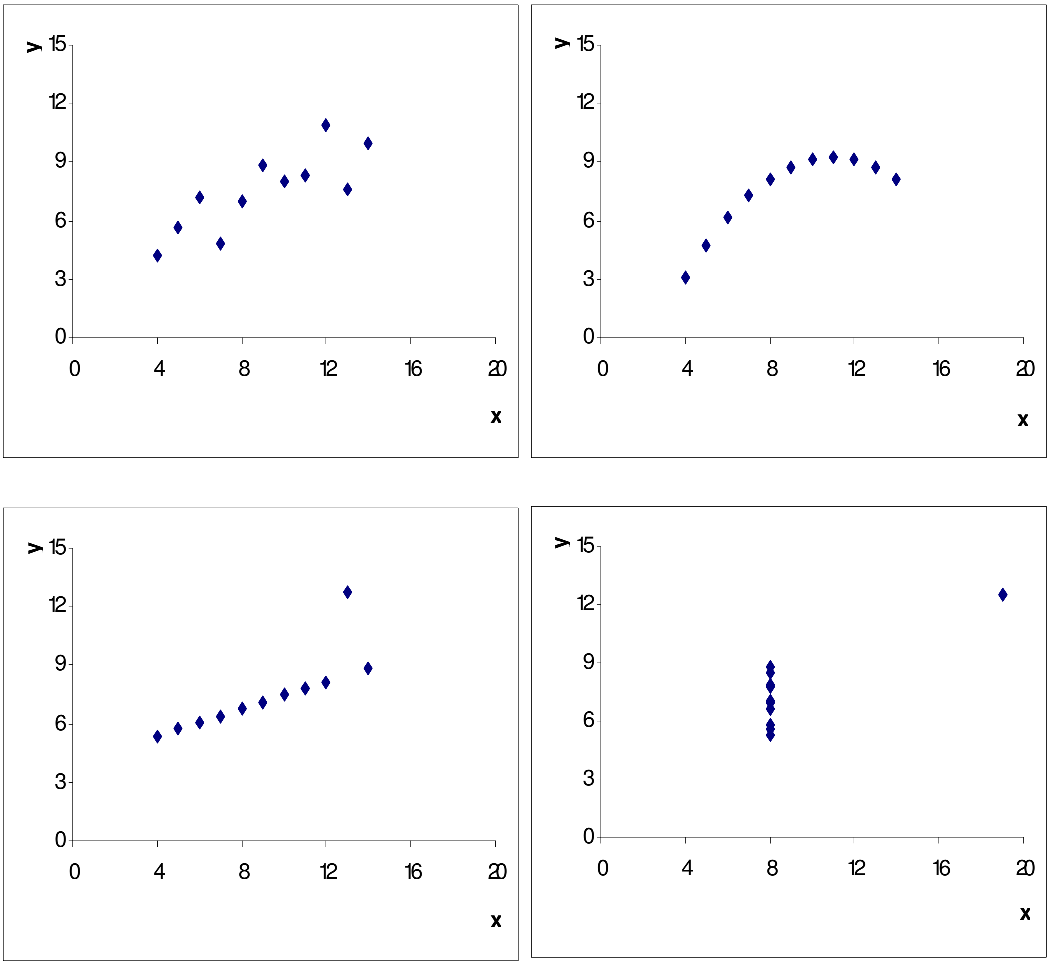


g. 14,5, 57.18, 77.1, 98.3 and 108.2; IQR = 41.12; no outliers since no observations beyond (-4.5, 159.98)
h. mean = 76.2851 and stdev = 22.32574; no outliers since no observations beyond (9.31, 143.26)

Solution Case 5.3

a. Straightforward.
b.

Figure. The scatter plots of Anscombe's datasets.



The upper-left plot shows a positive linear relation and the upper-right a complex relationship that certainly is non-linear. The lower-left plot pictures a perfect linear relation with one outlier. The lower-right plot demonstrates no variability in the *x*-data with the exception of an outlier in the upper right quadrant.

Solutions Cases Chapter 6

Solution Case 6.1 See book

Solution Case 6.2
A = Black; B = Blue; C = Green; D = Red

A beats B with probability 2/3;
B beats C with probability 2/3;
C beats D with probability 2/3.

And now the big surprise: D beats A with probability 2/3.

So let your opponent pick any die, and you know which die to choose in order to (most likely) beat him or her in a game of rolls, where the rules are as follows:

1 Highest roll scores a point
2 The player who reaches ten points first wins the game

Proof that C beats D two out of three times; D = Red = r and C = green = g:

D							
5	r	r	r	r	g	g	
5	r	r	r	r	g	g	
5	r	r	r	r	g	g	
1	g	g	g	g	g	g	
1	g	g	g	g	g	g	
1	g	g	g	g	g	g	
	2	2	2	2	6	6	C

Solutions Cases Chapter 7

Solution Case 7.1 See book

Solution Case 7.2

Intuitive solution:

a. From A to D in 2 steps: A→C→D; prob. = 0.1×0.1 = 0.01
b. From B to B in 2 steps: B→A→B or B→B→B; prob. = 0.1×0.2 + 0.7×0.7 = 0.51
c. From A to C in 2 steps: A→A→C or A→B→C or A→C→C; Prob. = 0.08 + 0.01 + 0.09 = 0.18
d. The person will ever reach D and then stay there.

Formal solution:

Write A_1, B_2, etc for the event that A, B, etc is reached of 1, 2 steps.

a. From A: $P(D_2) = P(C_1 \cap D_2) = P(C_1) \times P(D_2 \mid C_1)$ = 0.1×0.1 = 0.01
b. From B: $P(B_2) = P(A_1 B_2) + P(B_1 B_2)$
   $= P(A_1) \times P(B_2 \mid A_1) + P(B_1) \times P(B_2 \mid B_1)$
   = 0.1×0.2 + 0.7×0.7 = 0.51
c. From C: $P(C_2) = P(A_1 C_2) + P(B_1 C_2) + P(C_1 C_2)$
   = 0.8×0.1 = 0.1×0.1 + 0.1×0.9 = 0.18
d. See above.

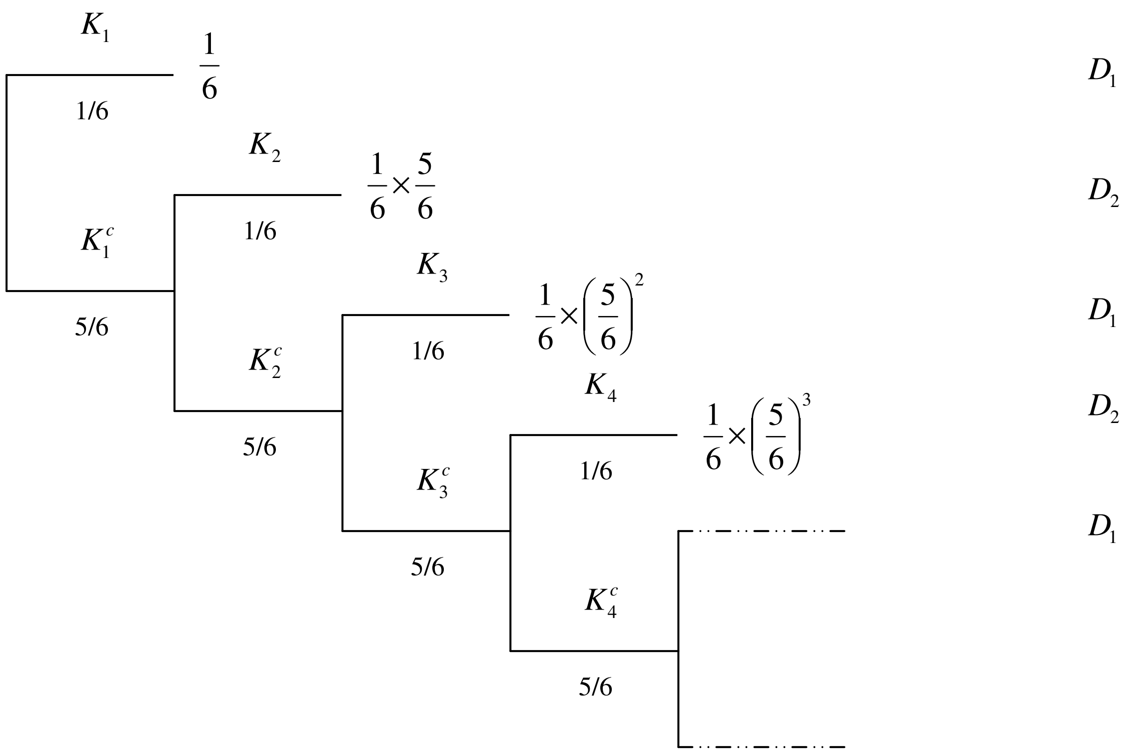
Solution Case 7.3

Intuitively it seems to be a disadvantage to start up the game. Below, the survival probability of the beginner will be calculated with the help of a probability tree.

Suppose that the game is played by the players 1 and 2, and that 1 starts the game. Consider the events:

K_i: the i^{th} shot is fatal; $i = 1, 2, \cdots$
D_j: player j looses; $j = 1, 2$

Notice that interest is in the probability $P(D_1)$. Since the successive shots are done independently, the following probability tree describes the game.



Some of the paths in the tree lead to D_1; they are indicated on the right-hand side. It follows that:

$$
\begin{aligned}
P(D_1) &= P(K_1) + P(K_1^c \cap K_2^c \cap K_3) + P(K_1^c \cap \cdots \cap K_4^c \cap K_5) + \cdots \\
&= \frac{1}{6} + \frac{1}{6}\times\left(\frac{5}{6}\right)^2 + \frac{1}{6}\times\left(\frac{5}{6}\right)^4 + \cdots
\end{aligned}
$$

Notice that the last expression is an ongoing summation of terms. The first term is a = 1/6 and – from the second term onwards – the terms arise from their predecessors by multiplication with $r = (5/6)^2$. From mathematical theory it is known that such ongoing summations are equal to a / (1 – r). Hence,

$$P(D_1) = \frac{1/6}{1-(5/6)^2} = \frac{6}{11}$$

Indeed, the probability that the player who starts the game will finally get the worst is larger than 0.5.

Solution case 7.4

Let A be the event that the overall experiment of randomly drawing ten teams results in five matches between a strong team and a weak team. Below, the probability $P(A)$ that A occurs will be calculated, while assuming that the draw is fair. Two (of many possible) solutions are presented.

Solution 1 (using the classical definition of probability and counting-rules) The overall experiment can be considered to be the result of ten consecutive sub-experiments: the consecutive drawings of the ten teams from the bowl. At each individual drawing, all remaining teams have the same probability of being selected. Hence the classical definition of probability is applicable for each sub-experiment, so that it suffices to count (at each drawing) the total number of outcomes as well as the number of outcomes leading to A (denoted as 'A-outcomes'). See the table below.

Drawing no.	1	2	3	4	5	6	7	8	9	10
total # outcomes	10	9	8	7	6	5	4	3	2	1
# *A*-outcomes	10	5	8	4	6	3	4	2	2	1

The second row hardly needs an explanation, since after each drawing, one team disappears from the bowl. For the first two numbers of the third row, we have the following arguments:

Drawing no. 1 : any team can lead to A,
2 : the team necessarily must belong to the opposite group.

So, if the first drawing results in a strong team, the second team has to be a weak one and vice versa. This corresponds with 10 and 5 possibilities, respectively. After these two drawings in accordance with event A, each group consists of four teams. Then the same reasoning leads to 8 and 4 possibilities for drawing no. 3 and 4, respectively. We continue in this way.

For the overall experiment with sample space Ω, we obtain the total number of outcomes N and the number of outcomes $N(A)$ favouring the event A, by applying the multiplication counting-rule (see Section 7.2):

$$N = 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 10!$$
$$N(A) = 10 \times 5 \times 8 \times 4 \times 6 \times 3 \times 4 \times 2 \times 2 \times 1 = 2^5 \times (5!)^2.$$

The classical definition leads to

$$P(A) = N(A) \,/\, N = 2^5 \times (5!)^2 \,/\, 10! = \frac{8}{63} = 0.1270.$$

Solution 2 (with conditional probabilities) Recall that A is the event that the ten drawings lead to five matches between a strong and a weak team. For $i = 1, 2, \cdots, 10$, let A_i denote the event that the i^{th} drawing is in accordance with A. Notice that

$$A = A_1 \cap A_2 \cap \ldots \cap A_{10}.$$

We can use the (generalised) product rule to rewrite $P(A)$ as follows:

$$P(A_1) \times P(A_2 \mid A_1) \times P(A_3 \mid A_1 \cap A_2) \times \cdots \times P(A_{10} \mid A_1 \cap A_2 \cap \ldots \cap A_9).$$

Notice that $P(A_1) = 1$, since every outcome of the first drawing agrees with A. For $P(A_2 \mid A_1)$, note that in the second drawing 5 out of 9 teams are in accordance with event A; hence this probability equals 5/9 (since the drawings are assumed to be fair). If we already know that drawings 1 and 2 both are in agreement with A (that is, $A_1 \cap A_2$ has occurred), then one of the matches is determined and all remaining 8 teams may lead to A. Hence,

$$P(A_3 \mid A_1 \cap A_2) = 1.$$

If it is already given that the first three drawings are all in accordance with event A, then 7 teams are left for the fourth drawing and 4 of them agree event A. Consequently,

$$P(A_4 \mid A_1 \cap A_2 \cap A_3) = 4/7.$$

We continue in this way. In the end, it follows that

$$P(A) = 1 \times \frac{5}{9} \times 1 \times \frac{4}{7} \times 1 \times \frac{3}{5} \times 1 \times \frac{2}{3} \times 1 = \frac{8}{63} = 0.1270.$$

Comments. The conclusion is that, under fair circumstances, the probability that the five strong countries are paired with the five weak countries is 0.1270. The outcome of the 'UEFA Euro 2004 play-offs draw' is remarkable, but can hardly be classified as suspicious. As a comparison: if a fair die is thrown, the probability of getting 6 is 0.167, only slightly more. And nobody will call an outcome 6 suspicious.

Apparently, the claim that $P(A)$ equals 0.0313 is incorrect. Probably without knowing, the people who adopted this claim intuitively assumed that the events $A_1, \cdots, A_{10}$ (for which the relation $A = A_1 \cap A_2 \cap \ldots \cap A_{10}$ holds) are independent and that $P(A_i) = 1$ for all odd i and $P(A_i) = 1/2$ for even i. But this assumption obviously is not valid.

Solutions Cases Chapter 8

Solution Case 8.1 See book

Solution Case 8.2
Reformulation of the facts of Case 6.3 in terms of X and probability yields:

Q1: 1. $P(X = 10000) = 1$
2. $P(X = 50000) = 0.1$; $P(X = 10000) = 0.5$; $P(X = 0) = 0.4$

Q2: 1. $P(X = 10000) = 1$
2. $P(X = 50000) = 0.15$; $P(X = 10000) = 0.45$; $P(X = 0) = 0.4$

Q3: 1. $P(X = 50000) = 0.2$ and $P(X = 0) = 0.8$
2. $P(X = 20000) = 0.5$ and $P(X = 0) = 0.5$

a. Q1: 1. $E(X) = 10000$ and $SD(X) = 0$
2. $E(X) = 50000 \times 0.1 + 10000 \times 0.5 + 0 \times 0.4 = 10000$

$$\begin{aligned} V(X) &= E(X^2) - 10000^2 \\ &= 50000^2 \times 0.1 + 10000^2 \times 0.5 + 0^2 \times 0.4 - 10000^2 \\ &= 200000000 \end{aligned}$$

$SD(X) = 14142.1356$

Q2: 1. $E(X) = 10000$ and $SD(X) = 0$
2. $E(X) = 12000$ and $SD(X) = 17888.5438$

Q3: 1. $E(X) = 10000$ and $SD(X) = 20000$
2. $E(X) = 10000$ and $SD(X) = 10000$

b. For Q1, option 1 will be chosen since it has the same expected payoff but no risk at all.
For Q2, the choice remains partially personal. Those who are not too afraid to take a risk will choose option 2.
For Q3, option 2 will be chosen since it has the same expected payoff but the standard deviation is smaller.

c. For Q1: respective utilities are 10000 and 8585.7864; choose 1.
For Q2: respective utilities are 10000 and 10211.1456; choose 2.
For Q3: respective utilities are 8000 and 9000; choose 2.

d. Overall choice: Option 2 mentioned in Q2.

Solution case 8.3

a. $E(R_1) = 0.06$;

$$E(R_2) = 0.11 \times \frac{1}{3} + 0.02 \times \frac{1}{3} + 0.05 \times \frac{1}{3} = 0.06;$$

$$E(R_3) = 0.01 \times \frac{1}{3} + 0.11 \times \frac{1}{3} + 0.21 \times \frac{1}{3} = 0.11$$

$V(R_1) = 0$;

$$V(R_2) = \frac{1}{3} \times (0.11 - 0.06)^2 + \frac{1}{3} \times (0.02 - 0.06)^2 + \frac{1}{3} \times (0.05 - 0.06)^2 = 0.0014;$$

$$V(R_3) = \frac{1}{3} \times (0.01 - 0.11)^2 + \frac{1}{3} \times (0.11 - 0.11)^2 + \frac{1}{3} \times (0.21 - 0.11)^2 = 0.0067$$

Assets 1 and 2 have the same expected return. Since asset 1 is riskless, it will always be preferred above asset 2. The expected return of asset 3 is much larger than the expected return of asset 1, but the volatility of 3 is also larger than the volatility of 1.

b. $u_1 = 0.06 - 5 \times 0 = 0.06$;

$u_2 = 0.06 - 5 \times 0.0014 = 0.0530$;

$u_3 = 0.11 - 5 \times 0.0067 = 0.0765$

An investor with $\alpha = 10$ will never prefer investing money in asset 2 to investing money in asset 1. Asset 3 will be preferred to both assets 1 and 2.

c. $u_1 = 0.06 - 10 \times 0 = 0.06$;

$u_2 = 0.06 - 10 \times 0.0014 = 0.0460$;

$u_3 = 0.11 - 10 \times 0.0067 = 0.0430$

An investor with $\alpha = 20$ will prefer asset 1 to both assets 2 and 3.

d. $E(R) = E(0.1R_1) + E(0.5R_2) + E(0.4R_3)$

$= 0.1\,E(R_1) + 0.5\,E(R_2) + 0.4\,E(R_3) = 0.08$

The table below gives the possible outcomes of R:

	recessive (*r*)	**neutral (*n*)**	**expansive (*e*)**
return portfolio	0.065	0.060	0.115

For instance, the outcome 0.065 of R follows from the multiplication 0.1×0.06 + 0.5×0.11 + 0.4×0.01.

Hence,

$$V(R) = E(R^2) - 0.08^2 = 0.065^2 \times \frac{1}{3} + 0.060^2 \times \frac{1}{3} + 0.115^2 \times \frac{1}{3} - 0.08^2$$

$= 0.007017 - 0.0064 = 0.000617$.

Notice that the variance for this portfolio is smaller than all three individual variances.

e. $u = 0.08 - 5 \times 0.000617 = 0.0769$, which is larger than the utilities of the individual assets.

f. $u = 0.08 - 10 \times 0.000617 = 0.0738$, which is larger than the utilities of the individual assets.

Solutions Cases Chapter 9

Solution Case 9.1 See book

Solution Case 9.2

a. $X \sim Bin(300,\ p)$ where p= 0.765 is the population proportion of private mobile phone owners phoning **prepaid**. Note that $E(X)$ = 229.5 and $V(X)$ = 53.9325. Below, Z is a standard normal rv.

b. **Exact**: $P(X < 225) = 0.2460$
(binomdist(224,300,0.765,1))
Approximated: $P(X < 225) \approx P(Z < -0.6808) = 0.2480$
(normsdist(-0.6808), with continuity correction)

c. **Exact**: $P(230 \le X \le 250) = P(X \le 250) - P(X \le 229)$
$= 0.9985 - 0.4952 = 0.5033$
Approximated: $P(230 \le X \le 250) = P(X \le 250.5) -$
$P(X \le 229.5)$
$\approx 0.9979 - 0.5 = 0.4979$

d. **prepaid total**: $8.74X$ **Expected**: 8.74×229.5 = 2005.83
subscribers total: 91.84×(300 – X) **Expected**: 91.84×70.5 = 6474.72
expected total: 8480.55 euro

e. The pre-payers spend only 100×(2005.83/8480.55) = 23.65% of the total amount spent.

Solutions Cases Chapter 10

Solution Case 10.1 See book

Solution Case 10.2

a.

Height class (cm)	Number of Reds	Number of Greens	Ratio red/green
< 140	337	159	2.117
140 –< 150	7861	4502	1.746
150 –< 160	72559	50138	1.447
160 –< 170	263822	219454	1.202
170 –< 180	381169	381169	1.000
180 –< 190	219454	263822	0.832
190 –< 200	50138	72559	0.691
200 –< 210	4502	7861	0.573
≥ 210	159	337	0.472
Total	1000000	1000000	-------

b. The national basketball team will have about two times as many Greens as Reds if only height plays a role. This, of course, has nothing to do will other qualities, especially not with basketball talent.

Solutions Cases Chapter 11

Solution Case 11.1 See book

Solution Case 11.2

a. The cross table gives the joint frequencies of the two stocks for the 49 weeks:

Ah \ Ph	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	Total
-8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
-7	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
-6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-4	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	2
-3	0	0	0	1	0	0	0	1	1	0	0	1	0	0	0	0	4
-2	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	2
-1	1	1	0	1	0	1	1	1	0	2	0	1	0	0	0	0	9
0	0	0	0	0	0	1	0	1	0	2	0	0	1	0	0	0	5
1	0	1	0	1	1	2	1	1	1	1	0	0	0	1	0	0	10
2	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	2
3	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	2
4	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	3
5	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	3
6	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	2
7	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	2
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Total	1	2	0	3	1	5	3	8	6	9	5	2	2	1	0	1	49

The joint pdf arises by dividing all numbers in the interior part of the table by 49.

b.

	Ph	**Ah**
mean	0.5239	0.6718
var	9.4218	11.4175
stdev	3.0695	3.3790
cov	2.9718	
correl	0.2865	

c. Use the rules for linear combinations:

	Expected return	**Risk**
$R_1 = R_{ph}$	0.5239	3.0695
$R_2 = R_{ah}$	0.6718	3.3790
$R_3 = R_{rf} = 0.06$	0.0600	0
$R_4 = 0.5R_{ph} + 0.5R_{ah}$	0.5979	2.5876
$R_5 = 0.25R_{rf} + 0.25T_{ph} + 0.50R_{ah}$	0.4819	2.0460
$R_6 = 0.25R_{rf} + 0.50T_{ph} + 0.25R_{ah}$	0.4449	1.9524
$R_7 = R_{rf}/3 + R_{ph}/3 + R_{ah}/3$	0.4186	1.7251
$R_8 = ..R_{rf} + ..R_{ph} + ..R_{ah}$	……	……

d. …………………….

Solution Case 11.3 (worked out)

a. For the variance of R_p, the covariances $\sigma_{i,j}$ of the individual returns R_i and R_j are needed. Since R_1 is degenerated at 0.06, the covariances $\sigma_{1,2}$ and $\sigma_{1,3}$ are both 0. For $\sigma_{2,3}$ we obtain:

$$\begin{aligned}\sigma_{2,3} &= (0.11 - 0.06)(0.01 - 0.11)/3 + 0 + (0.05 - 0.06)(0.21 - 0.11)/3 \\ &= -0.0020\end{aligned}$$

b. Note that $\mu_p = E(R_p)$ and $\sigma_p^2 = V(R_p)$ satisfy:

$$\begin{aligned}E(R_p) &= w_1E(R_1) + w_2E(R_2) + w_3E(R_3) = 0.06w_1 + 0.06w_2 + 0.11w_3 \\ &= 0.06(1 - w_1 - w_2) + 0.06w_2 + 0.11w_3 = 0.06 + 0.05w_3 \\ V(R_p) &= w_1^2V(R_1) + w_2^2V(R_2) + w_3^2V(R_3) + 2w_1w_2\sigma_{1,2} + 2w_1w_3\sigma_{1,3} + 2w_2w_3\sigma_{2,3} \\ &= 0w_1^2 + 0.0014w_2^2 + 0.0067w_3^2 + 2w_1w_2 \times 0 + 2w_1w_3 \times 0 - 0.0040w_2w_3 \\ &= 0.0014w_2^2 + 0.0067w_3^2 - 0.0040w_2w_3\end{aligned}$$

c. For the utility of the portfolio we get:

$$\begin{aligned}U(\mu_p, \sigma_p^2) &= \mu_p - \frac{1}{2}\alpha\sigma_p^2 \\ &= 0.06 + 0.05w_3 - 0.5\alpha(0.0014w_2^2 + 0.0067w_3^2 - 0.0040w_2w_3)\end{aligned}$$

d. Notice that only two of the three weights are left. To find the optimal weights, the partial derivatives are put equal to 0:

$$\frac{\partial}{\partial w_2}U(\mu_p, \sigma_p^2) = 0 \quad \Leftrightarrow \quad 0 - 0.5\alpha(2 \times 0.0014w_2 - 0.0040w_3) = 0$$
$$\frac{\partial}{\partial w_3}U(\mu_p, \sigma_p^2) = 0 \quad \Leftrightarrow \quad 005 - 0.5\alpha(2 \times 0.0067w_3 - 0.0040w_2) = 0$$

This system of two equations with two unknowns can equivalently be written in matrix notation:

$$\alpha\begin{pmatrix} 0.0014 & -0.0020 \\ -0.0020 & 0.0067 \end{pmatrix}\begin{pmatrix} w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0.05 \end{pmatrix}$$

It has the following solution:

$$\begin{pmatrix} w_2 \\ w_3 \end{pmatrix} = \alpha^{-1}\begin{pmatrix} 0.0014 & -0.0020 \\ -0.0020 & 0.0067 \end{pmatrix}^{-1}\begin{pmatrix} 0 \\ 0.05 \end{pmatrix} = \alpha^{-1}\begin{pmatrix} 18.5874 \\ 13.0112 \end{pmatrix}$$

(Notice that the 2×2 matrix is just the matrix that contains all covariances $\sigma_{i,j}$ for i, j = 2, 3.) Indeed, this choice for w_2 and w_3 maximises the utility function since $\alpha > 0$. Since $w_1 = 1 - w_2 - w_3$, the optimal weights are obtained.

e. Investor C, with risk aversion coefficient $\alpha = 50$, finds the following weights for the optimal portfolio:

$$w_2 = 0.37; \quad w_3 = 0.26; \quad w_1 = 0.37$$

Although an investor would never choose asset 2 instead of asset 1 (same expected return, but more risk), asset 2 **is** part of the optimal portfolio. Investor C puts 37% of his money in asset 2 and 26% in asset 3; 37% is invested riskfree.

f. The triples of weights w_1, w_2, w_3 of the optimal portfolios for investors A and B are respectively:

$$-2.16, 1.86, 1.30 \quad \text{and} \quad -0.58, 0.93, 0.65$$

For the optimal portfolio, both investors have to borrow money from other investors (against an interest rate of 6%). It is said that they are ***short*** in the riskless asset.

g. Investor A:
Filling in into the equation of **c**. yields: 0.0925
Investor B:
Filling in into the equation of **c**. yields: 0.0763
Investor C:
Filling in into the equation of **c**. yields: 0.0628

Solutions Cases Chapter 12

Solution Case 12.1 See book

Solution Case 12.2

a. The population proportion p is of interest for all dummy variables (the proportion of the ones) and for the levels of EDU. The other variables are quantitative, so population means are of interest.

b. The sample proportion $\hat{P}$ will be used for the dummy variables and for the levels of EDU; for the quantitative variables the sample mean $\overline{X}$ will be used.

c.

variable	DS	DH	DP	DF	EDU1	EDU2	EDU3	EDU4	EDU5
$\hat{p}$	0.237	0.763	0	0.173	0.24	0.26	0.34	0.11	0.05

variable	WEIGHT	LENGTH	AGE	WAGE	HOURS	NKIDS
$\overline{x}$	76.6083	176.5853	40.2556	5.8890	30.877	0.85
s	11.2890	8.8521	14.9722	4.8451	19.1238	1.129
variable	FS	FINC	FOODEXP	HOUSEXP	CLOTEXP	RECREXP
$\overline{x}$	2.69	30.6623	8.3579	11.2977	2.5283	2.5755
s	1.377	23.7717	4.8698	9.1032	1.8633	2.0458

d. The estimates of the population standard deviations are included in the table of **c**.

e. The mean education level of the population is estimated to be 2.47, with accompanying variance $1.1190^2 = 1.2522$.

f. **Correlations**

		FINC	FOODEXP	HOUSEXP	CLOTEXP	RECREXP
FINC	Pearson Correlation	1	.974(**)	.988(**)	.996(**)	.995(**)
	Sig. (2-tailed)		.000	.000	.000	.000
	N	300	300	300	300	300
FOODEXP	Pearson Correlation	.974(**)	1	.955(**)	.979(**)	.969(**)
	Sig. (2-tailed)	.000		.000	.000	.000
	N	300	300	300	300	300
HOUSEXP	Pearson Correlation	.988(**)	.955(**)	1	.980(**)	.986(**)
	Sig. (2-tailed)	.000	.000		.000	.000
	N	300	300	300	300	300
CLOTEXP	Pearson Correlation	.996(**)	.979(**)	.980(**)	1	.988(**)
	Sig. (2-tailed)	.000	.000	.000		.000
	N	300	300	300	300	300
RECREXP	Pearson Correlation	.995(**)	.969(**)	.986(**)	.988(**)	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	300	300	300	300	300

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

FINC is very strongly correlated to all expenditure variables.

Solutions Cases Chapter 13

Solution Case 13.1 See book

Solution Case 13.2

a. Thanks to the Central Limit Theorem, the 12 variables $\dfrac{\overline{X}-\mu}{\sigma_{\overline{X}}}$ are all approximately $N(0, 1)$ distributed. Hence:

$$\begin{aligned} P(\overline{X}-2\sigma_{\overline{X}}<\mu<\overline{X}+2\sigma_{\overline{X}}) &= P(\mu-2\sigma_{\overline{X}}<\overline{X}<\mu+2\sigma_{\overline{X}}) \\ &= P(-2<\frac{\overline{X}-\mu}{\sigma_{\overline{X}}}<2) \\ &= 2P(Z\le 2)-1 = 0.9545 \ (*) \end{aligned}$$

Note that this is the answer for all twelve samples. Apparently, it is likely (with 95.45% probability) that the 12 population means will fall in the 12 random intervals $(\overline{X}-2\sigma_{\overline{X}}, \overline{X}+2\sigma_{\overline{X}})$.

b. Since $\sigma_{\overline{X}}=\sigma/\sqrt{300}$, the 12 standard deviations $\sigma_{\overline{X}}$ follow immediately from the 12 population standard deviations that are given. Since the 12 sample means were calculated in Case 12.2, these results can be used to find the realisations of $(\overline{X}-2\sigma_{\overline{X}}, \overline{X}+2\sigma_{\overline{X}})$. The results are in the table:

variable	WEIGHT	LENGTH	AGE	WAGE	HOURS	NKIDS
$\overline{x}$	76.6083	176.5853	40.2556	5.8890	30.877	0.85
σ	12.02	9.41	15.10	6.11	19.70	1.18
$\overline{x}-2\sigma_{\overline{X}}$	75.2203	175.4987	38.5120	5.1835	28.6022	0.7137
$\overline{x}+2\sigma_{\overline{X}}$	77.9963	177.6719	41.9992	6.5945	33.1518	0.9863
variable	FS	FINC	FOODEXP	HOUSEXP	CLOTEXP	RECREXP
$\overline{x}$	2.69	30.6623	8.3579	11.2977	2.5283	2.5755
σ	1.35	15.59	3.78	5.89	1.29	1.33
$\overline{x}-2\sigma_{\overline{X}}$	2.5341	28.8621	7.9214	10.6176	2.3793	2.4219
$\overline{x}+2\sigma_{\overline{X}}$	2.8459	32.4625	8.7944	11.9778	2.6773	2.7291

For instance, it is very likely (95% certainty) that the present mean annual household income (the population mean) will lie between €28862.10 and €32462.50.

Solutions Cases Chapter 14

Solution Case 14.1 See book

Solution Case 14.2

a. Interest is in the four population proportions of ones of the variables DS, DH, DP and DF, and in the five population proportions of the levels of EDU. Note that the random sample of 300 households yields estimators $\hat{P}$ for each of these population proportions. Thanks to the Central Limit Theorem, the 9 variables $\frac{\hat{P}-p}{\sigma_{\hat{P}}}$ are all approximately $N(0, 1)$ distributed. Hence:

$$\begin{aligned} P(\hat{P}-2\sigma_{\hat{P}} < p < \hat{P}+2\sigma_{\hat{P}}) &= P(p-2\sigma_{\hat{P}} < \hat{P} < p+2\sigma_{\hat{P}}) \\ &= P(-2 < \frac{\hat{P}-p}{\sigma_{\hat{P}}} < 2) \\ &= 2P(Z \le 2) - 1 = 0.9545 \ (*) \end{aligned}$$

Note that this is the answer for all nine population proportions. Apparently, it is likely (with 95.45% probability) that the 9 population proportions will fall in the corresponding 9 random intervals $(\hat{P}-2\sigma_{\hat{P}}, \hat{P}+2\sigma_{\hat{P}})$.

b. Recall that $\sigma_{\hat{P}} = \sqrt{p(1-p)/300}$. Since the proportions p are unknown, the standard deviations $\sigma_{\hat{P}}$ can – in contrast to $\sigma_{\overline{X}}$ of Case 13.2 – not be observed when the data are known.

c. Replacement of p in $\sigma_{\hat{P}} = \sqrt{p(1-p)/300}$ by $\hat{P}$ yields $\sqrt{\hat{P}(1-\hat{P})/300}$. Since it is expected that $\hat{P}$ is close to p, the consequences for the probabilities in **a.** are that:

$$P(\hat{P}-2\sqrt{\hat{P}(1-\hat{P})/300} < p < \hat{P}+2\sqrt{\hat{P}(1-\hat{P})/300}) \approx 0.95$$

The population proportions p are probably included between

$$\hat{P}-2\sqrt{\hat{P}(1-\hat{P})/300} \quad \text{and} \quad \hat{P}+2\sqrt{\hat{P}(1-\hat{P})/300}$$

d.

variable	DS	DH	DP	DF	EDU1	EDU2	EDU3	EDU4	EDU5
$\hat{p}$	0.237	0.763	0	0.173	0.24	0.26	0.34	0.11	0.05
estimate of $\sigma_{\hat{P}}$	0.0246	0.0246	0	0.0218	0.0247	0.0253	0.0273	0.0181	0.0126
$\hat{p}-2\sqrt{\hat{p}(1-\hat{p})/300}$	0.1879	0.7139	0	0.1293	0.1907	0.2094	0.2853	0.0739	0.0248
$\hat{p}+2\sqrt{\hat{p}(1-\hat{p})/300}$	0.2861	0.8121	0	0.2167	0.2893	0.3106	0.3947	0.1461	0.0752

For instance, it is very likely (95% certainty) that the population proportion of households with female heads will lie between 0.1293 and 0.2167. The population proportion of households with level 5 educated heads, probably lies between 0.0248 and 0.0752. Recall that the population proportion with DP = 1 is known to be equal to 0.

Solutions Cases Chapter 15

Solution Case 15.1 See book

Solution Case 15.2
The table summarises the dataset:

	O		C		E		A		N	
	mean	size	mean	size	mean	size	mean	size	mean	size
1	0.0334	62	-0.0039	62	-0.2924	62	-0.1967	62	0.1637	62
2	-0.1378	64	0.0816	64	-0.1101	64	0.2636	64	0.0032	64
3	0.7186	14	-0.3519	14	0.3125	14	0.1396	14	0.3077	14
4	-0.0242	128	0.0125	128	0.1737	128	-0.0543	128	-0.1271	128
male	-0.0202	122	-0.0947	122	0.1345	122	-0.2494	122	-0.2773	122
female	0.0184	146	0.0904	146	-0.1026	146	0.2062	146	0.2207	146

Below, the tests suggested in the text of this case will be conducted. Here, irrespective of the trait that is considered, μ_i denotes the population mean for graduates of stream i; μ_m and μ_f denote the population means for the male and the female graduates. Note that the accompanying population variances are assumed to be equal to the overall population variance (which is 1). For all tests, the test statistics have the form:

$$\frac{\overline{X} - 0}{1/\sqrt{n}} = \sqrt{n}\overline{X}$$

Is μ_4 for E positive? Since ***val*** = $\sqrt{128} \times 0.1737 = 1.9652$ has the p-value 0.0247, the answer is Yes at significance level 0.05.
Is μ_1 for C positive? Since ***val*** = $\sqrt{62} \times -0.0039 = -0.0307$ has the p-value 0.5122 (!!), there is no evidence that Yes is the answer (at significance level 0.05).
Is $\mu_3 \neq 0$ for O? Since ***val*** = $\sqrt{14} \times 0.7186 = 2.6888$ has the p-value 0.0072 (two-sided), the answer is Yes at significance level 0.05.
Is μ_2 for A positive? Since ***val*** = $\sqrt{64} \times 0.2636 = 2.1088$ has the p-value 0.0175, the answer is Yes at significance level 0.05.
Is μ_m for N negative? Since ***val*** = $\sqrt{122} \times -0.2773 = -3.0629$ has the p-value 0.0011, the answer is Yes at significance level 0.05.

Solutions Cases Chapter 16

Solution Case 16.1 See book

Solution Case 16.2
From the dataset, the following sample sizes, means and standard deviations are measured:

variable	EMPFT	EMPPT	NMGRS	PSODA	PFRY	PENTREE
n	398	400	404	388	382	386
mean	8.2751	18.6775	3.4839	1.0449	0.9412	1.3541
standard deviation	7.97076	10.69964	1.13990	0.09357	0.10930	0.64970

For testing whether the population means of the variables EMPFT, EMPPT and NMGRS have changed, the respective values of the test statistics of the t-tests are 0.1880, –0.2851 and 1.1267. The accompanying p-values of the two-sided tests are 0.8510, 0.7757 and 0.2605. Since these p-values are all above the usually used significance levels, the conclusion is that there is no evidence that these means have changed.

For the variable EMPTOT = EMPFT + EMPPT + NMGRS, the population mean before 1 April 1992, was 30.45 (the sum of the three individual means). After this date, the sample mean and standard deviation of the 396 restaurants (that recorded all three variables) are 30.3485 and 12.42450. For testing whether the accompanying population mean is smaller than 30.45, the value of the test statistic of the t-test is –0.1626 and the (one-sided) p-value is 0.4355. The data do not support the statement that the mean of the total number of employees per restaurant has decreased.

For testing whether the population means of the three price variables PSODA, PFRY and PENTREE have changed, the respective values of the test statistics of the t-tests are 1.0315, 3.7909 and 1.0312. The accompanying p-values of the two-sided tests are 0.3030, 0.0002 and 0.3031. The conclusion is that there **is** evidence that the price of small fries has changed (increased).

The final conclusion is that an effect on the numbers of employees could not be detected. However, the price increase of small fries might be caused by the increase in minimum wage.

Solutions Cases Chapter 17

Solution Case 17.1 See book

Solution Case 17.2
For both the return data and the range data, it is assumed that the 50 observations are typical for the "restless" period after 25-07-2007; that they are the realisations of random samples.

a. It is additionally assumed that the daily returns (%) of the FTSE 100 index are normally distributed. We want to show that the population standard deviation σ of the returns in that restless period is larger than 0.90; or (equivalently) that $\sigma^2 > 0.81$. It is this (alternative) hypothesis that will be tested.

   From the data it follows easily that the sample standard deviation is 1.455808. For the ***val*** and the p-value we obtain:

   $$\boldsymbol{val} = \frac{49 \times (1.455808)^2}{0.81} = 128.2092;$$
   $$p\text{-value} = P(W \geq 128.2092) = 5.2 \times 10^{-9}$$

   The conclusion is that the standard deviation of the returns in that period after 25-07-2007 was larger than the "usual" standard deviation.

b. Let μ denote the mean of the intraday ranges of the FTSE 100 price (max – min) during the restless period after 25-07-2007. It is the alternative hypothesis H_1: $\mu > 50$ that will be tested. From the data in the second column it follows that the sample mean is 82.9980 with standard deviation 38.846345. For the ***val*** and the p-value we obtain:

   $$\boldsymbol{val} = \frac{82.9980 - 50}{38.846345/\sqrt{50}} = 6.0065;$$
   $$p\text{-value} = P(T \geq 6.0065) = 1.14 \times 10^{-7} \quad (*)$$

   The conclusion is that the intraday volatility of the price of the FTSE 100 index has increased after 25-07-2005.

Both conclusions reflect that volatility has (temporarily) increased.

Solutions Cases Chapter 18

Solution Case 18.1 See book

Solution Case 18.2

a. Note that the two samples are **paired** by way of date. That is why we use the paired-samples t-test that is based on the difference of the fund-returns and the corresponding AEX-returns.
If 1 refers to the fund and 2 to the AEX-index, then the testing problem is:

$$H_0\text{: } \mu_1 - \mu_2 \le 0 \text{ vs } H_1\text{: } \mu_1 - \mu_2 > 0$$

Since $\overline{d} = 0.00000823$, $s_D = 0.833345$ and $n = 510$, we obtain (df = 509):

$$\boldsymbol{val} = \frac{0.00000823}{0.833345/\sqrt{510}} = 0.0002230;$$

$$p\text{-value} = P(T \ge 0.0002230) = 0.4999 \ (*)$$

The data do not give evidence that the fund does on average better than the AEX.

b. Here is the five-step procedure:

(i) test H_0: $\dfrac{\sigma_1^2}{\sigma_2^2} \ge 1$ against H_1: $\dfrac{\sigma_1^2}{\sigma_2^2} < 1$; $\alpha = 0.05$

(ii) test statistic: $F = \dfrac{S_1^2}{S_2^2}$

(iii) reject $H_0 \iff f \le F_{0.95;252,256} = 0.8131$ (*)

(iv) $\boldsymbol{val} = \dfrac{0.670829}{0.910106} = 0.7371$

(v) reject H_0 since $\boldsymbol{val}$ belongs to the rejection region

The risk of the fund is smaller than the risk of the AEX-index.

Solution Case 18.3

Here is the test for 1: Netherlands and 2: Finland.

(i) H_0: $p_1 - p_2 = 0$ against H_1: $p_1 - p_2 \ne 0$ (so, hinge = 0)

(ii) test statistic: $Z = \dfrac{\hat{P}_1 - \hat{P}_2}{\sqrt{\dfrac{\hat{P}(1-\hat{P})}{n_1} + \dfrac{\hat{P}(1-\hat{P})}{n_2}}}$

(iii) $\hat{p} = \dfrac{316+392}{930+890} = 0.3890$

$$\boldsymbol{val} = \frac{0.34-0.44}{\sqrt{0.389\times 0.611\times(1/930+1/890)}} = -4.3743$$

(iv) p-value = $P(|Z| \ge |-4.3743|) = 2P(Z \ge 4.3743) = 0.0000122$

(v) reject H_0; the proportions are different

Solution Case 18.4

Note that a paired-samples t-test has to be conducted for **six** pairs of variables. The p-values are respectively: 0.952, 0.657, 0.244, 0.624, 0.000 and 0.122. As in Case 16.2, the conclusion is that only an effect on the price of small fries is detected.

Solutions Cases Chapter 19

Solution Case 19.1 See book

Solution Case 19.2

a. From the computer printout it follows that:

regression line: $\hat{y} = 32.639 + 0.074x$;

s_ε = 1866.6678; s_{B_1} = 0.00239; r^2 = 0.492;

val of t-test for significance of 'revenues', is 31.100.

From this two-sided t-test with hinge 0, it follows that the model is useful. This is also illustrated by the coefficient of determination: 49.2% of the variation in the profits is explained by the variation in the revenues.

b. If the revenues are one million dollars more, then the profit will on average be 0.074 million dollars more (which is 74000 dollars). We cannot interpret the intercept 32.639 since $x = 0$ is not part of the range of the revenues data.

c. The question is about the slope of the line of means, about being smaller than 0.08. So, the testing problem is: H_0: $\beta_1 \geq 0.08$ against H_1: $\beta_1 < 0.08$.

Since the ***val*** of the standard test is $\dfrac{0.074 - 0.08}{0.00239}$ = –2.5105 and the accompanying p-value is $P(T \leq -2.5105)$ = 0.0061(*), it can be concluded that the data do give evidence that $\beta_1 < 0.08$.

Solution Case 19.3

a. Regression of HRWAGEL on EDUCL (with n = 162) yields the estimated slope 2.556 and accompanying standard error 0.546. When testing whether β_1 > 0.5, the value of the test statistic is 3.7656 with p-value 0.0001.

Regression of HRWAGEH on EDUCH (with n = 161) yields the estimated slope 1.270 and accompanying standard error 0.373. When testing whether β_1 > 0.5, the value of the test statistic is 2.0643 with p-value 0.0203.

In both cases it is – at significance level 0.05 – concluded that one extra year of education on average increases hourly wage by more than \$0.50.

b. Some outliers seem to seriously disturb normality, in both cases. The scatter plots of residuals on $\hat{y}$ do not show obvious heteroskedasticity. The scatter plots of residuals on EDUC do not show an obvious misspecification.

c. Regression of HRWAGEL – HRWAGEH on EDUCL – EDUCH (with n = 149) yields the estimated slope 1.478 with standard error 0.466. When testing whether β_1 > 0.7, the value of the test statistic is 1.6695 with p-value 0.0486.

Regression of HRWAGEL – HRWAGEH on the difference between the cross-reported educations of twin 1 and twin 2 (with n = 149) yields the estimated slope 1.540 with standard error 0.451. When testing whether β_1 > 0.7, the value of the test statistic is 1.8625 with p-value 0.0323.

In both cases it is concluded at significance level 0.05 that one extra year for difference in education will on average yield more than \$0.70 extra difference in hourly wage.

e. It is concluded at significance level 0.05 that one extra year of difference in education for two people with similar family backgrounds will on average cause more than \$0.70 extra difference in hourly wage.

Solutions Cases Chapter 20

Solution Case 20.1

Solution Case 20.2

a. Regression of HRWAGEL on AGE and EDUCL (with n = 162) yields the respective estimated regression coefficients 0.241 and 2.640 with standard errors 0.105 and 0.541. Both variables are significant at level 0.05; $r^2 = 0.149$.
   Regression of HRWAGEH on AGE and EDUCH (with n = 161) yields the respective estimated regression coefficients 0.217 and 1.441 with standard errors 0.075 and 0.369. Both variables are significant at level 0.05; $r^2 = 0.115$.
   In both cases it is (at significant level 0.05) concluded that one extra year of education ceteris paribus and on average increases hourly wage by more than \$0.60.
b. Regression of HRWAGEL – HRWAGEH on AGE, EDUCL – EDUCH, DTENU and DUNCOVE (with n = 147) yields the following respective estimated regression coefficients (the accompanying standard errors are between brackets):

   0.213 (0.084), 1.410 (0.446), 0.474 (0.120), 1.711 (1.807)

   Only DUNCOVE is insignificant at level 0.05; $r^2 = 0.191$.
   Regression of HRWAGEL – HRWAGEH on AGE, DEDUCxx, DTENU and DUNCOVE (with n = 147) yields the following respective estimated regression coefficients (the accompanying standard errors are between brackets):

   0.215 (0.083), 1.562 (0.431), 0.481 (0.118), 2.008 (1.795)

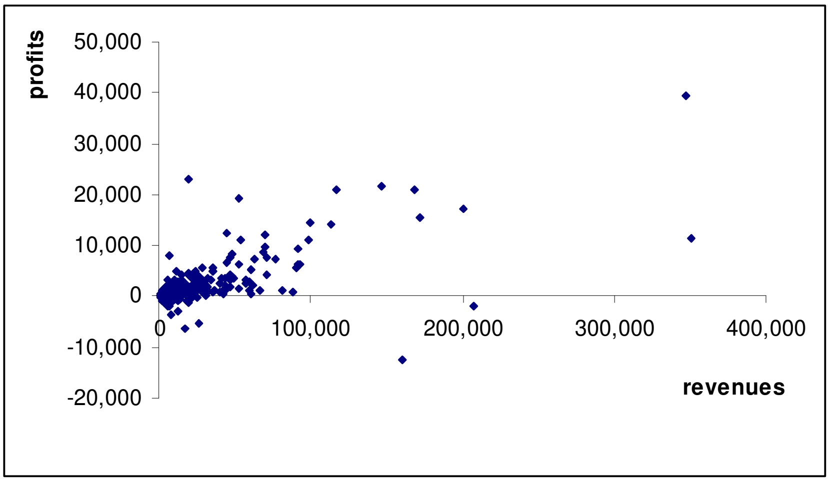
   Again, only DUNCOVE is insignificant at level 0.05; $r^2 = 0.207$.
   In both cases it is (because of the ***val***s 1.8161 and 2.2320) concluded at significance level 0.05 that one extra year of difference in education ceteris paribus and on average yields more than \$0.60 difference in hourly wage.
c. The estimated effect on hourly wage of difference in education is larger for the cross-reported educations, as was to be expected.
d. It is concluded at significance level 0.05 that one extra year of difference in education for one of two people with similar family and working backgrounds will on average increase the corresponding difference in hourly wage by more than \$0.60.

Solutions Cases Chapter 21

Solution Case 21.1 See book

Solution Case 21.2

a.



b. $E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

c. **Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.709(a)	.502	.501	1849.63190

a Predictors: (Constant), Rev3, Revenues, Rev2

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3440267910.122	3	1146755970.041	335.197	.000(a)
	Residual	3407453594.610	996	3421138.147		
	Total	6847721504.732	999			

a Predictors: (Constant), Rev3, Revenues, Rev2
b Dependent Variable: Profits

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-194.977	81.714		-2.386	.017
	Revenues	.110	.008	1.039	12.987	.000
	Rev2	-3.77E-007	.000	-.877	-3.810	.000
	Rev3	7.55E-013	.000	.565	3.267	.001

a Dependent Variable: Profits

The regression equation follows immediately from the coefficients part of the printout. From the model F-test it is concluded that the model is useful, with $r^2 = 0.502$. By t-tests it follows that the variables X, X^2 and X^3 are all individually significant within the model. Note that the coefficients of X^2 and X^3 are not far from 0. However, the accompanying standard deviations are even closer to 0 (note that they are **not** equal to 0), so that both variables still are significant.

d. The 95%- CI turns out to be: (649.46, 889.23)

Solution Case 21.3
(In this analysis, the PM data are included.)
Linear regression model for two-factor anova:

Model 1: $E(Y) = \beta_0 + \beta_1 DF + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_3$

(Base levels for 'gender' and 'study stream': male (value 1) and marketing (value 4).)
The model has to be run five times, once for each trait. The table contains summarised results.

trait	useful? $\alpha = 0.10$	gender effect? $\alpha = 0.05$	stream effect? $\alpha = 0.05$	r^2	pos/neg significance $\alpha = 0.025$
O	yes	no	yes	0.032	$\beta_4 > 0$
C	no	no	no	----	----
E	yes	no	yes	0.052	$\beta_2 < 0$
A	yes	yes	no	0.070	$\beta_1 > 0$
N	yes	yes	no	0.074	$\beta_1 > 0$

Linear regression model for two-factor anova with interaction:

Model 2:

$$E(Y) = \beta_0 + \beta_1 DF + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_3 + \beta_5 DFD_1 + \beta_6 DFD_2 + \beta_7 DFD_3$$

With respect to this model, note that (for example):

$$\beta_3 = E(Y \mid FM; DF = 0) - E(Y \mid mark; DF = 0)$$

$$\beta_3 + \beta_6 = E(Y \mid FM; DF = 1) - E(Y \mid mark; DF = 1)$$

Hence, β_6 is just the difference between the two mean-score-differences for FM and marketing: the mean-scores differences for female and male graduates.

trait	useful?	r^2	significance
O	no	----	----
C	no	----	----
E	yes	0.097	$\beta_1 < 0$; $\beta_3 < 0$; $\beta_6 > 0$
A	yes	0.076	$\beta_1 > 0$
N	yes	0.113	$\beta_1 > 0$; $\beta_3 > 0$; $\beta_6 < 0$

The usefulness for O is a bit narrow, which explains why the without-interaction model is significant at the level 0.10 but the with-interaction model is not.

Openness: PM graduates tend to score higher than marketing graduates, although the significance of the models is doubtful (model 1 is significant at level 0.10, but model 2 is not).
Conscientiousness: With respect to this trait, no significant gender or study stream effect is detected.

Extraversion: A stream effect is present; marketing graduates are more extravert than FM graduates. From model 2 it follows that this difference in extraversion is mainly caused by female graduates (since $\beta_6 < 0$).

Agreeableness: A gender effect is present; female graduates score higher than male graduates.

Neuroticism: A gender effect is present; female graduates are more neurotic than male graduates. From model 2 it follows that this difference is mainly caused by marketing (since $\beta_6 < 0$).

Note that (interval) estimates are wanted of respectively:

$$E(Y \mid FM; DF = 0) = \beta_0 + \beta_3,$$
$$E(Y \mid FM; DF = 1) = \beta_0 + \beta_1 + \beta_3 + \beta_6,$$
$$E(Y \mid mark; DF = 0) = \beta_0,$$
$$E(Y \mid mark; DF = 1) = \beta_0 + \beta_1$$

Point-estimates follow from the coefficients-part of the printout of model 2:

0.194; $-$0.066; $-$0.437; 0.360

For 95% confidence intervals, four new cases have to be created and 95%-CIs have to be determined for the expectations. Here are the respective results:

(-0.2619, 0.6507), (-0.3404, 0.2084), (-0.6497, -0.2236), (0.0990, 0.6208)

It follows that the expected score for the male marketing graduate is significantly negative and that the expected score for the female marketing graduate is significantly positive. Both conclusions have confidence levels (at least) 0.95.

Solution Case 21.4

In this solution, Netherlands is chosen as base level. New model:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_5 + \beta_4 D_{FR} + \cdots + \beta_{10} D_{UK}$$

Five-step procedure (p-value approach) to test for a country-effect:

(i) H_0: $\beta_4 = \cdots = \beta_{10} = 0$ vs H_1: at least one of $\beta_4, \cdots, \beta_{10}$ is $\neq 0$

(ii) test statistic: $F = \dfrac{(SSE_r - SSE_c)/7}{SSE_c/(n-11)}$

(iii) ***val*** $= \dfrac{(43559901 - 35755518)/7}{35755518/368} = 11.4748$

(iv) p-value $= P(F \geq 11.4748) = 3.42 \times 10^{-13}$ (*)

(v) H_0 is rejected; there exists a country effect

At significance level 0.05, the coefficients β_4, β_6 are positive and β_5 is negative. That is: France and Italy have larger emission, Germany smaller.

X_1 and X_3 are still individually significant (again: positively, respectively negatively). Note that X_5 is not significant within this new model. Apparently, the inclusion of country dummies makes the factor 'wind' superfluous.

Solutions Cases Chapter 22

Solution Case 22.1 See book

Solution Case 22.2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.714(a)	.510	.506	1839.41582

a Predictors: (Constant), DTexas, DFemceo, Revenues, DCalifornia, DNewYork, Rev3, Rev2

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3491338534.596	7	498762647.799	147.412	.000(a)
	Residual	3356382970.136	992	3383450.575		
	Total	6847721504.732	999			

a Predictors: (Constant), DTexas, DFemceo, Revenues, DCalifornia, DNewYork, Rev3, Rev2
b Dependent Variable: Profits

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-288.854	89.382		-3.232	.001
	Revenues	.108	.008	1.015	12.700	.000
	Rev2	-3.67E-007	.000	-.852	-3.720	.000
	Rev3	7.46E-013	.000	.558	3.244	.001
	DFemceo	-127.327	373.369	-.008	-.341	.733
	DCalifornia	169.798	193.562	.020	.877	.381
	DNewYork	766.947	204.500	.086	3.750	.000
	DTexas	269.195	188.370	.032	1.429	.153

a Dependent Variable: Profits

It follows that the model is useful, with $r^2 = 0.510$. The gender dummy has p-value 0.733, which indicates that no significant difference exists between the mean for the profits of corporations with female CEOs and the mean for the profits of comparable corporations with male CEOs.

To find out whether a state effect is present, the model without the state-dummies has to be estimated too. Here is the ANOVA part of the printout:

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
2	Regression	3440341509.468	4	860085377.367	251.156	.000(a)
	Residual	3407379995.264	995	3424502.508		
	Total	6847721504.732	999			

a Predictors: (Constant), DFemceo, Rev3, Revenues, Rev2
b Dependent Variable: Profits

A partial F-test leads to the conclusion that there indeed is a state-effect (***val*** = 5.0242). From the printout of the first model it follows (by a one-sided t-test) that corporations in the state New York on average have larger profits than comparable corporations in states other than New York, California or Texas. A similar conclusion cannot be drawn for the state California, neither for the state Texas.

Solution Case 22.3

a. Basic assumption:

$$E(W) = \beta_0 + \beta_1 age + \beta_2 dedl2 + \beta_3 dedl3 + \beta_4 dedl4 + \beta_5 dedl5 + \beta_6 fem + \beta_7 age^2$$
$$+ \beta_8 age * dedl2 + \beta_9 age * dedl3 + \beta_{10} age * dedl4 + \beta_{11} age * dedl5$$

Here are the printouts of the complete model (with the interaction terms) and the reduced model (without the interaction terms):

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9544.043	11	867.640	19.286	.000(a)
	Residual	6208.254	138	44.987		
	Total	15752.297	149			

a Predictors: (Constant), AGE_DEDL5, AGE_DEDL4, FEM, AGE_DEDL2, AGE2, DEDL3, DEDL2, AGE_DEDL3, DEDL5, DEDL4, AGE
b Dependent Variable: W

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.341	6.588		.507	.613
	AGE	.689	.373	.735	1.850	.067
	DEDL2	-4.688	4.948	-.195	-.947	.345
	DEDL3	2.360	5.163	.110	.457	.648
	DEDL4	3.386	7.229	.123	.468	.640
	DEDL5	-24.852	9.683	-.658	-2.567	.011
	FEM	-3.000	1.160	-.146	-2.587	.011
	AGE2	-.006	.005	-.443	-1.074	.285
	AGE_DEDL2	.158	.150	.218	1.050	.296
	AGE_DEDL3	.055	.151	.090	.367	.714
	AGE_DEDL4	.073	.199	.099	.369	.713
	AGE_DEDL5	.995	.236	1.145	4.218	.000

a Dependent Variable: W

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
2	Regression	8588.704	7	1226.958	24.321	.000(a)
	Residual	7163.594	142	50.448		
	Total	15752.297	149			

a Predictors: (Constant), AGE2, DEDL3, FEM, DEDL5, DEDL4, DEDL2, AGE
b Dependent Variable: W

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
2	(Constant)	7.516	6.441		1.167	.245
	AGE	.295	.368	.315	.801	.425
	DEDL2	.301	1.879	.013	.160	.873
	DEDL3	4.521	1.797	.211	2.516	.013
	DEDL4	6.098	2.125	.222	2.869	.005
	DEDL5	15.786	2.602	.418	6.068	.000
	FEM	-3.122	1.219	-.152	-2.561	.011
	AGE2	.002	.005	.139	.358	.721

a Dependent Variable: W

The first model is useful, with $r^2 = 0.574$. Only DEDL5, FEM and AGE*DEDL5 are significant at level 0.05, while AGE is significant at level 0.10.

The partial F-test has ***val*** = 5.3089, which is larger than $F_{0.05;4,138} = 2.4373$ (*).

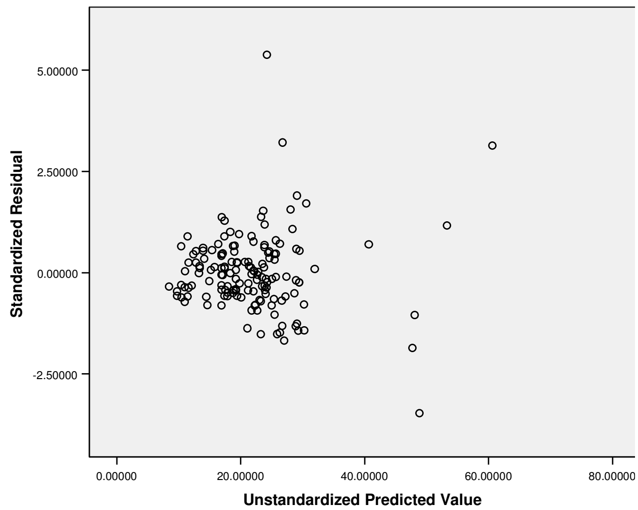
Hence, it is concluded that the interaction terms are useful within the model.

b. The (complete) model has several disadvantages: only three variables are significant at significance level 0.05. Furthermore, AGE and AGE2 are both individually insignificant, which is caused by collinearity. The inclusion of the interaction terms and AGE2 has caused that these terms seriously complicate the interpretation of the model.

Omitting the interaction terms makes the model statistically a bit worse, but it becomes easier to be interpreted: DEDL4, DEDL5 and FEM are individually significant.

Omitting AGE2 too makes interpretation even easier.

c. Here is the scatter plot of the standardised residuals on $\hat{w}$:



Note that the variation increases with increasing value of $\hat{w}$. The picture obviously shows heteroskedasticity.

d.

$$E(LW) = \beta_0 + \beta_1 lage + \beta_2 lage2 + \beta_3 dedl2 + \beta_4 dedl3 + \beta_5 dedl4 + \beta_6 dedl5 + \beta_7 fem$$

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
3	Regression	18.276	7	2.611	37.475	.000(a)
	Residual	9.893	142	.070		
	Total	28.169	149			

a Predictors: (Constant), DEDL5, DEDL4, FEM, DEDL2, LAGE2, DEDL3, LAGE
b Dependent Variable: LW

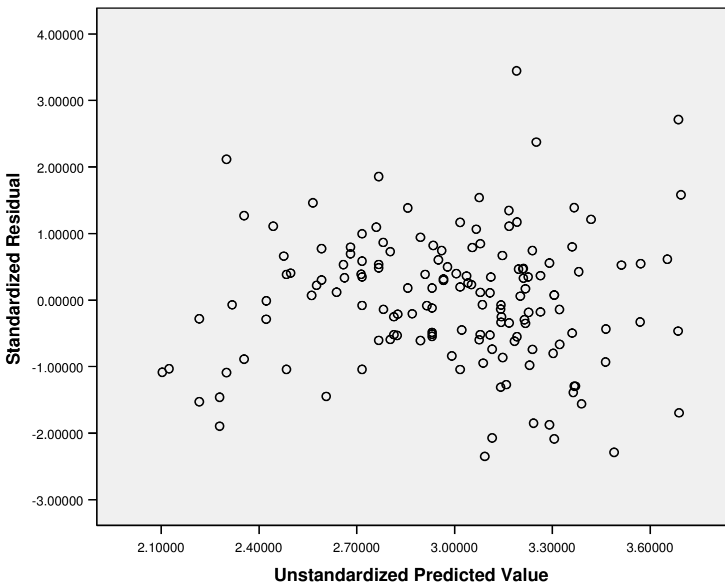
Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
3	(Constant)	-8.153	2.576		-3.165	.002
	LAGE	5.680	1.489	4.211	3.814	.000
	LAGE2	-.709	.213	-3.660	-3.331	.001
	FEM	-.122	.045	-.141	-2.690	.008
	DEDL2	-.021	.070	-.021	-.298	.766
	DEDL3	.155	.069	.171	2.258	.025
	DEDL4	.213	.081	.183	2.635	.009
	DEDL5	.476	.097	.298	4.892	.000

a Dependent Variable: LW

The model is useful, with $r^2 = 0.649$. Only the variable DEDL2 is not individually significant at level 0.05, which indicates that there is no evidence of a difference between the mean gross hourly wage of level 1 educated persons and the mean gross hourly wage of level 2 educated persons with the same gender and age. However, for levels 3, 4 and 5 the means of the hourly wages are significantly **larger** than the mean hourly wage of level 1 educated persons.

e.



The heteroskedasticity problem has reduced when compared to **c**.

f. Here is the partial printout for that extended model:

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
4	(Constant)	-7.488	2.686		-2.788	.006
	LAGE	5.365	1.532	3.978	3.501	.001
	LAGE2	-.674	.217	-3.477	-3.106	.002
	FEM	-.559	.497	-.645	-1.124	.263
	DEDL2	-.025	.070	-.025	-.358	.721
	DEDL3	.155	.069	.171	2.261	.025
	DEDL4	.208	.081	.179	2.561	.012
	DEDL5	.482	.098	.302	4.933	.000
	LAGE_FEM	.127	.143	.497	.883	.379

a Dependent Variable: LW

It turns out that this interaction term is not significant; there is no significant interaction.

g. Here is the printout for that extended model:

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
5	Regression	18.935	11	1.721	25.724	.000(a)
	Residual	9.234	138	.067		
	Total	28.169	149			

a Predictors: (Constant), LAGE_DEDL5, LAGE_DEDL4, FEM, LAGE_DEDL2, LAGE2, DEDL3, DEDL2, LAGE_DEDL3, DEDL5, DEDL4, LAGE
b Dependent Variable: LW

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
5	(Constant)	-9.340	2.809		-3.325	.001
	LAGE	6.503	1.650	4.822	3.941	.000
	LAGE2	-.847	.241	-4.372	-3.513	.001
	FEM	-.129	.045	-.149	-2.871	.005
	DEDL2	-1.002	.627	-.987	-1.597	.113
	DEDL3	.303	.679	.334	.447	.656
	DEDL4	-.179	.995	-.154	-.180	.857
	DEDL5	-2.681	1.315	-1.678	-2.039	.043
	LAGE_DEDL2	.291	.186	.969	1.569	.119
	LAGE_DEDL3	-.043	.197	-.165	-.219	.827
	LAGE_DEDL4	.114	.281	.348	.405	.686
	LAGE_DEDL5	.862	.358	2.006	2.405	.017

a Dependent Variable: LW

The ***val*** of the partial F-test turns out to be 2.4622, slightly larger than $F_{0.05;4,138}$ = 2.4373. Hence, at significance level 0.05 it can (justly) be concluded that the extension has some usefulness. But on the other hand, the individual significance of some of the education dummies is seriously disturbed.

h. The mean gross hourly wage of men is larger than the mean gross hourly wage of women even after having included AGE and the EDL-dummies in the model, as follows from the one-sided *t*-test; see the printout of **d**. The regression coefficient of FEM is –0.122, which is the estimated ceteris paribus difference between log(*W*) for women and men. If the man has hourly wage *w*, then the ceteris paribus woman is estimated to have her log-wage equal to log(w) – 0.122 and hence her wage equal to:

$$e^{\log(w)-0.122} = e^{\log(w)} \times e^{-0.122} = 0.885w$$

In the sample, the mean hourly wage of the women is 80% of the mean hourly wage of the men. If follows that 8.5% of this wage backlog is explained by differences in age and education. The other 11.5% is explained by variables that are not included in the model. These arguments give answers to the questions under 1.

According to the "final" model, the estimated ceteris paribus difference between log(W) for a level 5 educated person and a level 1 educated person, is 0482. If the level 1 educated person has hourly wage *w*, then the ceteris paribus level 5 person is estimated to have the log-wage equal to log(w) + 0.482 and hence W equal to:

$$e^{\log(w)+0.482} = e^{\log(w)} \times e^{0.482} = 1.62w$$

Hence, the mean hourly wage of level 5 educated persons is 62% more than the mean hourly wage of level 1 educated persons with the same age and gender.

Solutions Cases Chapter 23

Solution Case 23.1 See book

Solution Case 23.2 See book

Solutions Cases Chapter 24

Solution Case 24.1 See book

Solution Case 24.2

When running the standard χ^2-test with a computer package, it turns out that ***val*** = 218.774. Since the test uses 28 degrees of freedom and $\chi^2_{0.01;28}$ = 48.2782 (*), the conclusion is that there is evidence that the eight distributions are not all the same.

However, the printout also indicates that five cells have expected frequencies less than 5. Since these cells are all dealing with the values 4 and 5 of Quest 3 (as can be seen in a printout), we combine these values and conduct the χ^2-test (that now has 21 degrees of freedom) again. Since ***val*** = 195.269 and $\chi^2_{0.01;21}$ = 38.9322 (*), the conclusion is the same.

Solutions Cases Chapter 25

Solution Case 25.1

a.

b.

(i) test H_0: the 28 population locations are the same
against H_1: at least two population locations differ

(ii) test statistic: $W = \left[\frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{T_j^2}{n_j} \right] - 3(n+1)$

(iii) reject H_0 $\Leftrightarrow$ $w \geq \chi^2_{0.05;27} = 40.1133$ (*)

(iv) ***val*** = 1616.4

(v) the locations are not all the same

c, d. The table summarises the two sample proportions for 1, 2 and 3, 4:

country	**proportion 1, 2**	**proportion 3, 4**
Belgium	0.7715	0.2285
Czech Republic	0.6054	0.3946
Denmark	0.6651	0.3349
Germany	0.7724	0.2276
Estonia	0.6721	0.3279
Greece	0.9205	0.0795
Spain	0.8217	0.1783
France	0.8895	0.1105
Ireland	0.7090	0.2910
Italy	0.8836	0.1164
Cyprus	0.8665	0.1335
Latvia	0.9163	0.0837
Lithuania	0.8337	0.1663
Luxembourg	0.7959	0.2041
Hungary	0.8820	0.1180
Malta	0.8074	0.1926
Netherlands	0.6007	0.3993
Austria	0.6782	0.3218
Poland	0.8573	0.1427
Portugal	0.8962	0.1038
Slovenia	0.8660	0.1340
Slovakia	0.8781	0.1219
Finland	0.5991	0.4009
Sweden	0.7740	0.2260
United Kingdom	0.7305	0.2695
Norway	0.6709	0.3291
Iceland	0.5439	0.4561
United States	0.7181	0.2819

Largest: Greece; smallest: Finland