University of Waterloo STAT 221 – W. H. Cherry

Assignment 7 Outline Solution

A7 – 1. This question involves a χ^2 goodness-of-fit test of a probability model.

Data: The 62 observations (not actually given in the question) on the educational level of each grand juror.

Model: The empirical probability model is given in the question as the county distribution of educational levels by percentage in four categories (classes).

Freq.: The observed frequencies are given in the question; the expected frequencies are obtained from the county distribution of educational levels.

Educational level	Elementary	Secondary	Some college		College degree
Observed frequency	1	10	16	51	35
Expected frequency	17.608	30.070	7.378	14.322	6.944

NOTE: It is a serious mistake in this question to convert the *observed* frequencies to percentages and then to use these and the county percentages given in the question as the basis of the calculations which follow for the goodness-of-fit test. (Why?)

Dis. meas.:
$$\sum_{\text{classes}} \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}} = \frac{(1-17.608)^2}{17.608} + \ldots\ldots + \frac{(35-6.944)^2}{6.944}$$
$$= 15.6648 + 13.3956 + 10.0758 + 113.3553 \simeq 152.49.$$

NOTE: Although there is poor agreement between *all* the observed and expected frequencies in this problem, the major contribution to the value of the χ^2 statistic comes from the last educational class, *College degree*; such people appear to be greatly *over*represented among the 62 jurors.

D.f.: There are 4 frequency classes and no model parameters estimated from the data, so: $df = 4 - 0 - 1 = 3$.

***P*-value:** $\Pr[\chi^2_3 \geq 152.49] \simeq 7.65 \times 10^{-33}$

Answer: Using a χ^2 goodness-of-fit test of a probability model, there is (extremely) highly statistically significant evidence of lack of fit of the model ($P \simeq 7.65 \times 10^{-33}$); thus, we choose **option (b)** from among the five possibilities given in the question.

NOTE: If the last two frequency classes are combined, as indicated by the braces on the frequency table above, the value of the χ^2 statistic is 122.49 and the P-value is $\Pr[\chi^2_2 \geq 122.49] \simeq 2.52 \times 10^{-27}$; thus, the Answer is *not* meaningfully altered.

A7 – 2. This question involves a χ^2 goodness-of-fit test of a probability model.

Data: The observations (not actually given in the question) on the outcomes from the 360 rolls of the die.

Model: Using a 36-point sample space for one roll of two distinguishable fair dice, we can use counting arguments to find the probability distribution for the 11 possible outcomes as shown at the right.

Outcome	2	3	4	5	6	7	8	9	10	11	12
36×Probability	1	2	3	4	5	6	5	4	3	2	1

Freq.: The observed frequencies are given in the question, the expected frequencies come from multiplying the probabilities in the model by 360, the number of times the dice were rolled; thus, we have:

Outcome	2	3	4	5	6	7	8	9	10	11	12
Observed frequency	11	18	33	41	47	61	52	43	29	17	8
Expected frequency	10	20	30	40	50	60	50	40	30	20	10.

Dis. meas.:
$$\sum_{\text{classes}} \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}} = \frac{(11-10)^2}{10} + \frac{(18-20)^2}{20} + \ldots\ldots + \frac{(17-20)^2}{20} + \frac{(8-10)^2}{10}$$
$$= 0.1 + 0.2 + 0.3 + 0.025 + 0.18 + 0.01\dot{6} + 0.08 + 0.225 + 0.0\dot{3} + 0.45 + 0.4 = 2.01.$$

D.f.: There are 11 frequency classes and no model parameters estimated from the data, so: $df = 11 - 0 - 1 = 10$.

***P*-value:** $\Pr[\chi^2_{10} \geq 2.01] = 0.996\,263 \simeq 0.996$

Answer: Using a χ^2 goodness-of-fit test of a probability model, there is no statistically significant evidence of lack of fit of the model ($P \simeq 0.996$); thus, the data suggest that the dice **are** fair.

However, we might be concerned that the fit of the probability model is *too good*; *i.e.*, there may be some-

Assignment 7 Outline Solution (continued 1)

A7 – 2. Answer: (cont.) thing suspicious about how *close* the agreement is between the observed and expected frequencies – only about 4 times in 1,000 would such good agreement be expected for a pair of fair dice. We *might* therefore answer that it would **not** be a good idea to accept an offer to play a game of chance with the person using these two dice.

NOTE: The second part of the Answer for this question is reminiscent of the discussion in Program 24 of *Against All Odds: Inside Statistics* about Mendel's data showing *suspiciously* good agreement with his theory.

A7 – 3. This question involves a χ^2 goodness-of-fit test of a probability model.

Data: The observations (not actually given in the question) on the resistance to the brown plant hopper of each of the 374 lines of rice that were raised in the investigation.

Model: From the statement of the question, the IRRI model, under an assumption of *independence*, gives the three probabilities at the right.

Outcome	Resistant	Mixed	Susceptible
Probability	0.25	0.5	0.25

Freq.: The observed frequencies are given in the question, the expected frequencies come from multiplying the probabilities in the model by 374, the number of lines of rice that were raised; thus, we obtain the frequencies shown at the right.

Outcome	Resistant	Mixed	Susceptible
Observed frequency	97	184	93
Expected frequency	93½	187	93½

Dis. meas.: $$\sum_{\text{classes}} \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}} = \frac{(97-93.5)^2}{93.5} + \frac{(184-187)^2}{187} + \frac{(93-93.5)^2}{93.5}$$
$$= 0.1301 + 0.0481 + 0.0027 = 0.\dot{1}\dot{8}.$$

D.f.: There are 3 frequency classes and no model parameters estimated from the data, so: $df = 3 - 0 - 1 = 2$.

***P*-value:** $\Pr[\chi^2_2 \geq 0.\dot{1}\dot{8}] = 0.913\,101 \simeq 0.9$.

Answer: Using a χ^2 goodness-of-fit test of a probability model, there is no statistically significant evidence of lack of fit of the IRRI model ($P \simeq 0.9$); thus, the data **are** consistent with the assumption of (probabilistic?) independence used in the IRRI genetic model.

A7 – 4. (a) This question involves a χ^2 goodness-of-fit test of a probability model; because we can regard the *V1* flying bomb hits as *outcomes occurring unpredictably (on an individual basis) in (two-dimensional) space*, we use a *Poisson* model.

Data: The observations (not actually given in the question) on the number of hits in each of the ¼-square-kilometre areas.

Model: The Poisson model is: $\Pr(y \text{ hits in a given ¼-square-kilometre area}) = \dfrac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, 2, \ldots\ldots$

Without a value of the model parameter μ (the *mean* or *average rate* of the Poisson process) being given in the question, we must *estimate* it from the data as:

$$\text{estimate of } \mu = \frac{\text{total number of hits}}{\text{total number of areas}} = \frac{0\bullet229 + 1\bullet211 + \ldots\ldots + 7\bullet1}{576} = \frac{537}{576} \simeq 0.932\,292.$$

Using this value of μ, the relevant Poisson probabilities are:

Number of hits	0	1	2	3	4	5	6	7	≥8
Probability	0.393 651	0.366 997	0.171 074	0.053 164	0.012 391	0.002 310	0.000 359	0.000 048	0.000 006

Freq.: The observed frequencies are given in the table in the question and the expected frequencies are obtained by multiplying the Poisson probabilities tabulated above by 576, the number of ¼-square-kilometre areas; thus, we obtain the following frequencies:

Number of hits	0	1	2	3	4	5	6	7	≥8
Observed frequency	229	211	93	35	7	0	0	1	0
Expected frequency	226.74	211.39	98.54	30.62	7.14	1.33	0.21	0.03	0.00.

The expected frequencies in the last *four* classes are less than 5 which we take as the lower limit for an

Assignment 7 Outline Solution (continued 2)

A7 – 4. (cont.) (a) **Freq.:** acceptable approximation for the χ^2 distribution of the discrepancy measure; we therefore combine the classes as in the table at the right.

Number of hits	0	1	2	3	≥4
Observed frequency	229	211	93	35	8
Expected frequency	226.74	211.39	98.54	30.62	8.71

Dis. meas.: $\sum_{\text{classes}} \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}} = \frac{(229-226.74)^2}{226.74} + \ldots\ldots + \frac{(8-8.71)^2}{8.71}$

$= 0.02253 + 0.00072 + 0.31146 + 0.62653 + 0.05788 \simeq 1.01912.$

D.f.: There are 5 frequency classes and 1 model parameter estimated from the data, so: $df = 5 - 1 - 1 = 3.$

***P*-value:** $\Pr[\chi^2_3 \geq 1.01912] \simeq 0.796\,626 \simeq 0.8.$

Answer: Using a χ^2 goodness-of-fit test of a probability model, there is no statistically significant evidence of lack of fit of the model ($P \simeq 0.8$); thus, the distribution of the numbers of hits over the 576 ¼-square-kilometre areas appears to be closely in accordance with a Poisson model with a mean of about 0.932.

(b) As Feller (Volume 1, page 161) has pointed out, most people at the time believed in the tendency of the points of impact of the *V1*s to cluster; if this were true, there would be a higher frequency in the data of areas with many hits or no hits and a deficiency in the intermediate classes. Because of the close fit of a Poison model, the data in fact indicate essentially complete unpredictability (almost 'perfect randomness') and homogeneity over the whole region comprised of the 576 ¼-square-kilometre areas; this reminds us that unpredictability ('randomness') may be (wrongly) understood to imply an *absence* of clustering and that the *presence* of clustering can then be (wrongly) taken as evidence of predictability. (Of course, a *sufficient* degree of clustering *does* provide such evidence.)

A7 – 5. This question involves a χ^2 goodness-of-fit test in a 2×4 contingency table.

Hyp.: Hair colour and sex can be modelled as *probabilistically independent* in children.

Data The observations (*i.e.*, each child's hair colour and sex) for all the children in the nvestiagtion – these data are not actually given in the statement of the question.

Freq.: The *observed* frequencies (obtained from the data) are given in the table in the statement of the question. The *expected* frequencies are calculated (assuming the hypothesis of probabilistic independence is true) from the row and column totals in the table of observed frequencies, using the result:

expected frequency for the cell in row i and column j of a contingency table $= \frac{r_i \cdot c_j}{n}$;

they are shown in the table at the right.

SEX	HAIR COLOUR Fair	Red	Medium	Dark	
Boys	63.00	12.16	84.55	50.29	210
Girls	51.00	9.84	68.45	40.71	170
	114	22	153	91	380

Dis. meas.: $\sum_{\text{cells}} \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}} = \frac{(60-63.00)^2}{63.00} + \frac{(12-12.16)^2}{12.16} + \ldots\ldots + \frac{(68-68.45)^2}{68.45} + \frac{(38-40.71)^2}{40.71}$

$= 0.14286 + 0.00211 + 0.00240 + 0.14604$

$+ 0.17647 + 0.00260 + 0.00296 + 0.18040 \simeq 0.65582.$

D.f.: For a 2×4 table: $df = (2-1)(4-1) = 3.$

***P*-value:** $\Pr[\chi^2_3 \geq 0.65582] = 0.883\,542 \simeq 0.9.$

Answer: Using a χ^2 goodness-of-fit test 2×4 contingency table, the hypothesis of probabilistic independence is *not* rejected at the 5% (or even at the 75%) level ($P \simeq 0.9$); thus, the data provide no statistically significant evidence of an association between children's sex and their hair colour.

A7 – 6. This question involves a χ^2 goodness-of-fit test in a 3×4 contingency table.

(a) **Hyp.:** Students' smoking habits can be modelled as *probabilistically independent* of their parents' smoking habits.

Data The observations (*i.e.*, each student's smoking habits and those of their parents) for all the students in the investigation – these data are not actually given in the statement of the question.

(*continued overleaf*)

Assignment 7 Outline Solution (continued 3)

A7 – 6. (cont.) (a) **Freq.:** The *observed* frequencies (obtained from the data) are given in the table in the statement of the question.

The *expected* frequencies are calculated (assuming the hypothesis of independence is true) from the row and column totals in the table of observed frequencies, using the result:

expected frequency for the cell in row i and column j of a contingency table $= \frac{r_i \cdot c_j}{n}$;

they are shown in the table at the right.

STUDENTS' HABITS	PARENTS' HABITS Neither	Father	Mother	Both	
Non-smoker	102.08	64.38	38.86	26.68	232
Casual	51.04	32.19	19.43	13.34	116
Regular	22.88	14.43	8.71	5.98	52
	176	111	67	46	380

Dis. meas.: $\sum_{\text{cells}} \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}} = \frac{(115-102.08)^2}{102.08} + \frac{(70-64.38)^2}{64.38} + \ldots\ldots + \frac{(10-8.71)^2}{8.71} + \frac{(11-5.98)^2}{5.98}$

$= 1.6353 + 0.4906 + 2.5018 + 2.8239 + 1.9750 + 0.1490 + 3.7800$
$+ 1.0042 + 0.3625 + 0.8153 + 0.1911 + 4.2141 \simeq 19.9426.$

D.f.: For a 3×4 table: $df = (3-1)(4-1) = 6$.

***P*-value:** $\Pr[\chi^2_6 \geq 19.9426] = 0.002\,835 \simeq 0.003$.

Answer: Using a χ^2 goodness-of-fit test 3×4 contingency table, the hypothesis of probabilistic independence is rejected at the 1% level ($P \simeq 0.003$); thus, the data provide highly statistically significant evidence of an association between students' smoking habits and those of their parents.

To investigate the *form* of the association between the students' and their parents' smoking habits, we find the observed frequency in each cell as a *percentage* of its *column* total, as shown at the right. We see that the percentages *de*crease across the *first* row and generally *in*crease across the *second and third* rows; this indicates that children of *non*-smoking parents are *less* likely, and those of smoking parents are *more* likely, to smoke. The percentages also show a *dose response* in that the intermediate categories (*viz.* students who are *casual* smokers, only *one* smoking parent) usually show percentages *between* the values of the more extreme categories on either side.

STUDENTS' HABITS	PARENTS' HABITS Ne	Fa	Mo	Bo
Non-smoker	65	63	43	39
Casual	23	27	42	37
Regular	11	10	15	24

(b) Despite the association found in (a) and the reasonableness of its form with respect to what might be expected, the data can*not* be used to establish that children tend to follow their parents' smoking habits in a *causal* sense; the reason is that the Plan for this investigation is *observational* rather than *experimental* with *equiprobable assigning.* As Program 11 of *Against All Odds: Inside Statistics* reminded us, it is difficult to establish causation for the health and other consequences of tobacco use because, in investigations with human participants, it is both infeasible and unethical to assign smoking habits probabilistically; rather, such habits are *self-selected*.

A7 – 7. This question involves a χ^2 goodness-of-fit test in a 4×3 contingency table.

Hyp.: Level of alcohol consumption and level of nicotine consumption during pregnancy can be modelled as *probabilistically independent* of each other.

Data The observations (*i.e.*, each pregnant woman's level of alcohol consumption and nicotine consumption) for all the women in the investiagtion – these data are not actually given in the statement of the question.

Freq.: The *observed* frequencies (obtained from the data) are given in the table in the statement of the question.

The *expected* frequencies are calculated (assuming the hypothesis of probabilistic independence is true) from the row and column totals in the table of observed frequencies, using the result:

expected frequency for the cell in row i and column j of a contingency table $= \frac{r_i \cdot c_j}{n}$;

they are shown in the table at the right.

ALCOHOL (oz/day)	NICOTINE (mg/day) None	1-15	≥16	
None	82.73	17.69	22.59	123
.01-.10	51.12	10.93	13.96	76
.11-.99	109.63	23.44	29.93	163
≥1.00	60.53	12.94	16.53	90
	304	22	65	452

Dis. meas.: $\sum_{\text{cells}} \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}} = \frac{(105-82.73)^2}{82.73} + \frac{(7-17.69)^2}{17.69} + \ldots\ldots + \frac{(16-12.94)^2}{12.94} + \frac{(17-16.53)^2}{16.53}$

$= 5.9948 + 6.4599 + 5.9464 + 0.9259 + 3.2173 + 0.0660 +$
$5.9919 + 7.8444 + 4.8675 + 0.2059 + 0.7236 + 0.0134 \simeq 42.2571.$

Assignment 7 Outline Solution (continued 4)

A7 – 7. (cont.) **D.f.:** For a 3×4 table: $df = (3-1)(4-1) = 6$.

***P*-value:** $\Pr[\chi^2_6 \geq 42.2571] = 0.000\,000\,163\,587 \simeq 2 \times 10^{-7}$.

Answer: Using a χ^2 goodness-of-fit test in a 3×4 contingency table, the hypothesis of probabilistic independence is rejected at the 1% (and even at the 0.01%) level ($P \simeq 2 \times 10^{-7}$); thus, the data provide very highly statistically significant evidence of an association between levels of alcohol consumption and nicotine consumption among pregnant women.

To investigate the *form* of the association between the womens' levels of alcohol and nicotine consumption, we find the observed frequency in each cell as a *percentage* of its *column* total and of its *row* total, as shown in the two tables at the right. The generally *de*creasing percentages in the first two rows of the *upper* table and *in*creasing percentages in its last two rows indicate that there is a tendency towards greater alcohol consumption for women with higher nicotine consumption. The generally *de*creasing percentages down the first column of the *lower* table and *in*creasing percentages down the second and third columns indicate a tendency towards greater nicotine consumption for women with higher alcohol consumption. Thus, *both* tables show that the levels of consumption of alcohol and nicotine by pregnant women have a *positive* association – they tend to be low together and high together.

Percentages of column totals

ALCOHOL (oz/day)	NICOTINE (mg/day) None	1-15	≥16
None	35	11	13
.01-.10	19	8	16
.11-.99	28	57	51
≥1.00	19	25	20

Percentages of row totals

ALCOHOL (oz/day)	NICOTINE (mg/day) None	1-15	≥16
None	85	6	9
.01-.10	76	7	17
.11-.99	52	23	26
≥1.00	63	18	19

An *informal* assessment of the *strength* of the association is that it is of *moderate* strength; while the trends in the percentages in the tables at the right are clear, we see that they are usually *not* monotonic. Such 'non-ideal' behaviour is typical of real data sets.

Despite the association found by examining the two tables of percentages given above at the right, the data can*not* be used to establish that consumption of one of the two substances *causes* the consumption of the other, because these data come from an *observational* Plan, rather than from an experimental Plan involving *equiprobable assigning*. As Program 11 of *Against All Odds: Inside Statistics* reminded us, it is difficult to demonstrate causation in situations like this one where it would be infeasible and unethical to assign the level of either alcohol or nicotine consumption under equiprobable assigning; rather, such levels are *self-selected*.

A7 – 8. (a) This question involves a χ^2 goodness-of-fit test in a 3×3 contingency table.

Hyp.: The shape and colour classifications for the seeds can be modelled as *probabilistically independent*.

Data The observations (*i.e.*, each seed's shape and colour) for all the seeds in the investigation – these data are not actually given in the statement of the question.

Freq.: The *observed* frequencies (obtained from the data) are given in the statement of the question; we can set them out in a 3×3 contingency table as shown at the upper right, using the abbreviations $R \equiv$ *round* and $W \equiv$ *wrinkled* for seed *shape*, and $Y \equiv$ *yellow* and $G \equiv$ *green* for seed *colour*.

SEED SHAPE	SEED COLOUR *Y*	*YG*	*G*	
R	38	65	35	138
RW	60	138	67	265
W	28	68	30	126
	126	271	132	529

The *expected* frequencies are calculated (assuming the hypothesis of independence is true) from the row and column totals in the table of observed frequencies, using the result:

expected frequency for the cell in row i and column j of a contingency table $= \dfrac{r_i \cdot c_j}{n}$;

they are shown in the lower table at the right.

SEED SHAPE	SEED COLOUR *Y*	*YG*	*G*	
R	32.87	70.70	34.43	138
RW	63.12	135.76	66.12	265
W	30.01	64.55	31.44	126
	126	271	132	529

Dis. meas.: $\sum_{\text{cells}} \dfrac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}} = \dfrac{(38-32.87)^2}{32.87} + \dfrac{(65-70.70)^2}{70.70} + \ldots\ldots + \dfrac{(68-64.55)^2}{64.55} + \dfrac{(30-31.44)^2}{31.44}$

$= 0.80064 + 0.45955 + 0.00944 + 0.15422 + 0.03696 + 0.01171 + 0.13463 + 0.18439 + 0.06596 \simeq 1.85748.$

D.f.: For a 3×3 table: $df = (3-1)(3-1) = 4$.

(*continued overleaf*)

Assignment 7 Outline Solution (continued 5)

A7 – 8. (cont.) (a) ***P*-value:** $\Pr[\chi^2_4 \geq 1.85748] = 0.761\,950 \simeq 0.76$.

Answer: Using a χ^2 goodness-of-fit test 3×4 contingency table, the hypothesis of probabilistic independence is not rejected at the 1% (or even at the 50%) level ($P \simeq 0.76$); thus, the data provide no statistically significant evidence of an association between the colour and shape classifications of the seeds.

(b) This question involves a χ^2 goodness-of-fit test of a probability model.

Data: The observations (not actually given in the question) on the colour (yellow, yellow and green, green) of the 529 seeds studied in Mendel's experiment.

Model: The model provided by Mendel's theory of heredity predicts that the three seed colours should occur in the ratio 1:2:1, as stated in the question; this corresponds to the three probabilities shown at the right.

Outcome	*Y*	*YG*	*G*
Probability	0.25	0.5	0.25

Freq.: The observed frequencies are the *column totals* in the table of observed frequencies in (a) and the expected frequencies are obtained by multiplying the probabilities in the model by 529, the number of seeds examined in the investigation; thus, we obtain the frequencies shown above at the right.

Outcome	*Y*	*YG*	*G*	Total
Observed frequency	126	271	132	529
Expected frequency	132.25	264.5	132.25	529

Dis. meas.: $$\sum_{\text{classes}} \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}} = \frac{(126-132.25)^2}{132.25} + \frac{(271-264.5)^2}{264.5} + \frac{(132-132.25)^2}{132.25}$$
$$= 0.29537 + 0.15974 + 0.00047 \simeq 0.455\,5765.$$

D.f. There are 3 frequency classes and no model parameters estimated from the data, so: $df = 3 - 0 - 1 = 2$.

***P*-value:** $\Pr[\chi^2_2 \geq 0.455\,5765] = 0.796\,293 \simeq 0.8$.

Answer: Using a χ^2 goodness-of-fit test of a probability model, there is no statistically significant evidence of lack of fit of the model for seed colour based on Mendel's theory of heredity ($P \simeq 0.8$); thus, the data *are* consistent with Mendel's theory.

A7 – 9. (a) The description (in the question) of the data being obtained from *records* in 34 major hospitals makes it clear this investigation had an *observational* Plan – it found what happened to people under the conditions prevailing at the different hospitals, almost certainly with*out* equiprobable assigning in the choice of anesthetic (or blinding of the people who collected the data). Thus, the data in the question show *association* between anesthetic and the proportion of deaths among people to whom it is administered, but the *observational* nature of the Plan means we can*not* legitimately infer *causation* for *any* of the anesthetics and, specifically, we *cannot* infer that anesthetic C is *causing* a higher proportion of deaths than A, B or D.

(b) *Other* explanations, than anesthetic C being more dangerous, are possible for its association with a higher death rate; for example, C may have been used in higher-risk operations. To investigate this matter, additional information that should be sought from the hospital records includes:

- the type of operation performed on each person and its associated risk of death;
- patient characteristics such as sex, age and state of health prior to surgery;
- some measure of hospital size and death rate in surgery cases.

A7 – 10. (a) The scatter diagram of the Indianapolis winning speed data, with three estimated regressions of **Y** on **X** superimposed on it, is shown on the facing page 0.15g.

(b) From the five numerical summaries of the 55 data points given in the question, we can calculate the following:

$$\overline{\text{x}} = 1{,}948.3\dot{2}\dot{7}, \qquad \overline{\text{y}} = 124.314\,\dot{1}\dot{8},$$
$$\text{SS}_{\text{xy}} = 21{,}514.775, \qquad \text{SS}_{\text{x}} = 17{,}018.11, \qquad \text{SS}_{\text{y}} = 28{,}001.357\,95;$$

hence, the estimates of β_1 (the slope) and β_0 (the intercept) of the (model) regression of **Y** on **X** are:

$$b_1 = 1.264\,228\,225, \qquad b_0 = -2{,}338.816\,148,$$

so the equation of the straight-line model is: $_{\text{reg}}\overline{\text{y}} = -2{,}338.8 + 1.2642\text{x}$;

also, the coefficient of determination and the correlation coefficient are:

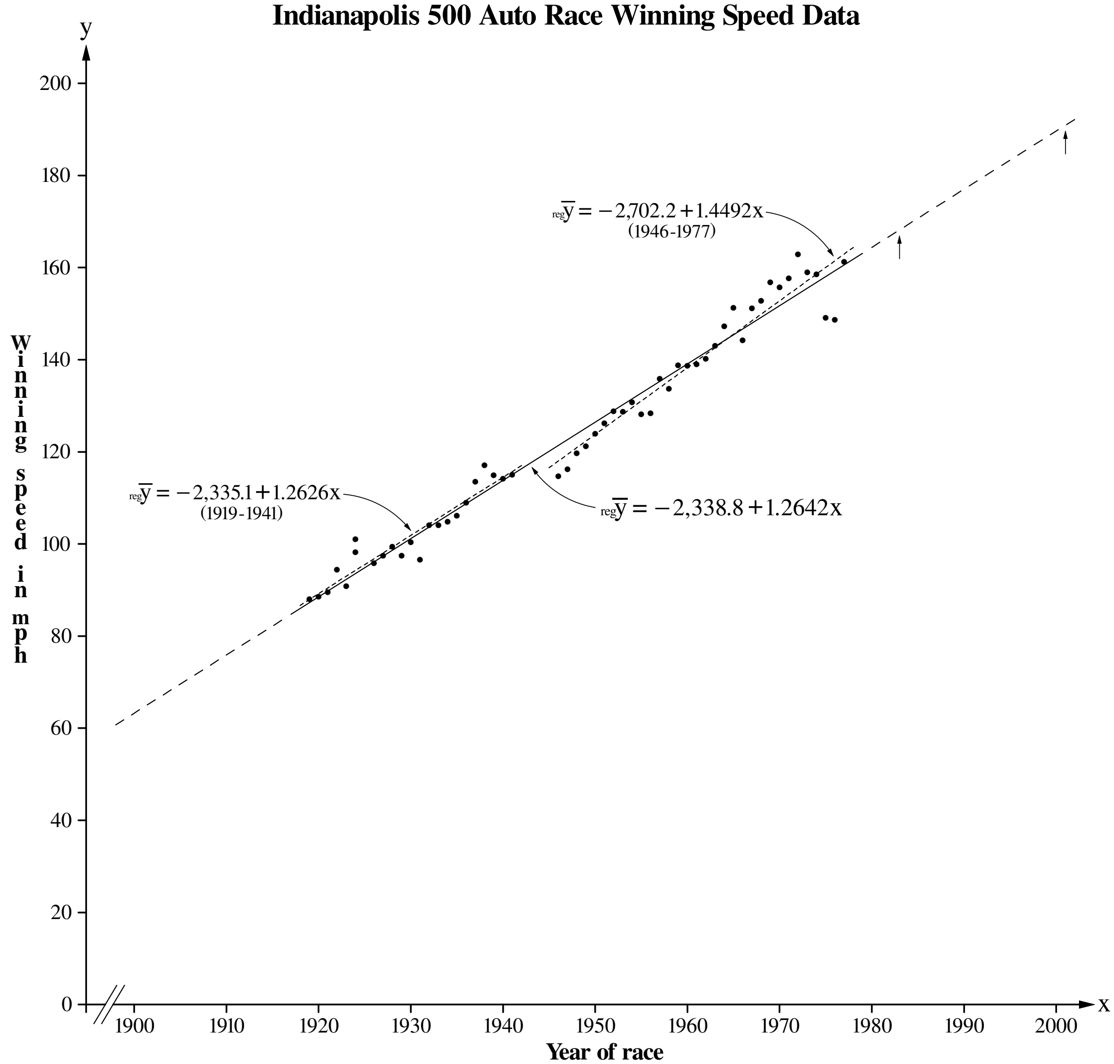
$$\text{r}^2 = 0.971\,366\,669, \qquad \text{r} = 0.985\,579.$$

(*continued*)

Assignment 7 Outline Solution (continued 6)

A7–10. (a)
(cont.)

Indianapolis 500 Auto Race Winning Speed Data



(c) We see from the scatter diagram, with the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$ superimposed on it, that the slope of the points (representing the *average rate of increase* of the winning speed of the race) is *greater* after the War years than before them. We could therefore improve the fit of a *straight line* model using *separate* lines for the pre- and post-war years (1919-1941 and 1946-1977 respectively).

NOTE: For interest, we can estimate the two *separate* straight-line regressions of $\mathbf{Y}$ on $\mathbf{X}$ as follows:

For 23 points: (1919-1941)	$\Sigma x_i = 44{,}390$	$\Sigma y_i = 2{,}342.29$	$\Sigma x_i y_i = 4{,}521{,}897.49$
	$\Sigma x_i^2 = 85{,}673{,}712$	$\Sigma y_i^2 = 240{,}313.2057$	$n = 23$;
	$\bar{x} = 1{,}930$,	$\bar{y} = 101.838\ 6957$,	
	$SS_{xy} = 1{,}277.79$,	$SS_x = 1{,}012$,	$SS_y = 1{,}777.447\ 27$;
	$b_1 = 1.262\ 638\ 340$,	$b_0 = -2{,}335.053\ 30$,	*i.e.*: $_{\text{reg}}\bar{y} = -2{,}335.1 + 1.2626x$;
	$r^2 = 0.909\ 683\ 481$,	$r = 0.953\ 773$.	
For 32 points: (1946-1977)	$\Sigma x_i = 62{,}768$	$\Sigma y_i = 4{,}494.99$	$\Sigma x_i y_i = 8{,}820{,}876.38$
	$\Sigma x_i^2 = 123{,}122{,}160$	$\Sigma y_i^2 = 637{,}659.0213$	$n = 32$;
	$\bar{x} = 1{,}961.5$,	$\bar{y} = 140.468\ 4375$,	
	$SS_{xy} = 3{,}953.495$,	$SS_x = 2{,}728$,	$SS_y = 6{,}254.799\ 43$;
	$b_1 = 1.449\ 228\ 372$,	$b_0 = -2{,}702.193\ 18$,	*i.e.*: $_{\text{reg}}\bar{y} = -2{,}702.2 + 1.4492x$;
	$r^2 = 0.916\ 019\ 320$,	$r = 0.957\ 089$.	

Both estimated regressions are shown dashed on the scatter diagram above.

(*continued overleaf*)

Assignment 7 Outline Solution (continued 7)

A7–10. (cont.) (d) We want a point estimate of the random variables $Y(\mathrm{x} = 1983)$ and $Y(\mathrm{x} = 2001)$; these are obtained by substituting the relevant value of x in the equation of the estimated regression line: $_{\text{reg}}\overline{\mathrm{y}} = -2{,}338.8 + 1.2642\mathrm{x}$. The results are:

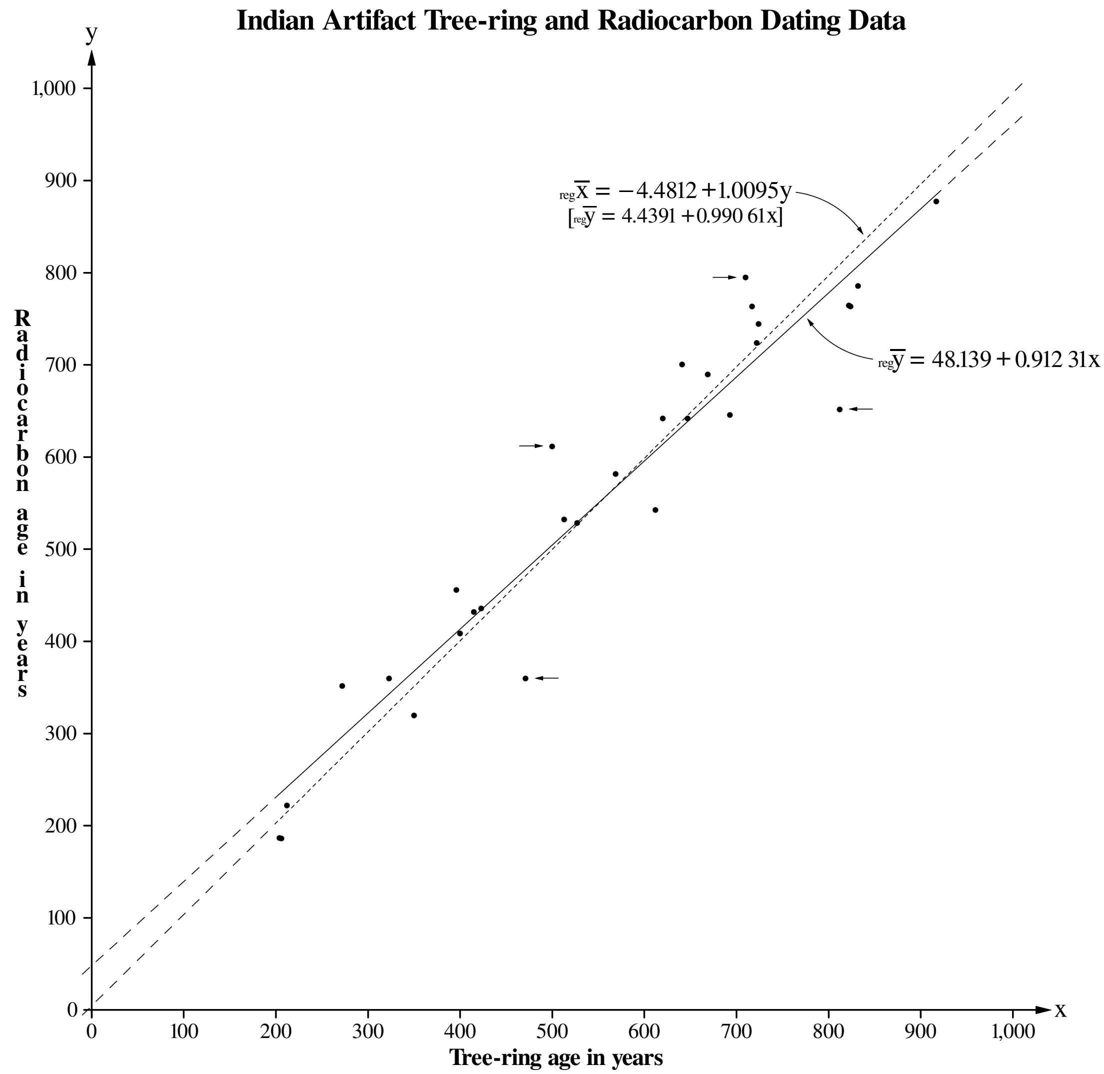
$\hat{\mu}_Y(\mathrm{x} = 1983) = 168.148$ mph or about 168 mph;

- this prediction should be fairly reliable because it is based on a model that is a fair fit to the data and the value of x is not too far outside the range of the observed values of x.

$\hat{\mu}_Y(\mathrm{x} = 2001) = 190.904$ mph or about 191 mph;

- this prediction is much less reliable because the value of x is appreciably *out*side the range of the observed values of x, and we have no information on whether the straight-line model applies for values of x around 2000.

A7–11. (a) The scatter diagram of the tree-ring dating data, with the estimated regressions of **Y** on **X** and **X** on **Y** superimposed on it, is:



(b) From the five numerical summaries of the 30 data points given in the question, we can calculate the following:

$\overline{\mathrm{x}} = 558.1$, $\overline{\mathrm{y}} = 557.3$,

$SS_{xy} = 1{,}110{,}301.1$, $SS_x = 1{,}217{,}020.7$, $SS_y = 1{,}099{,}878.3$;

hence, the estimates of β_1 (the slope) and β_0 (the intercept) of the (model) regression of **Y** on **X** are:

$b_1 = 0.912\,310\,776$, $b_0 = 48.139\,355\,56$,

so the estimated regression of **Y** on **X** is: $_{\text{reg}}\overline{\mathrm{y}} = 48.1394 + 0.912\,311\mathrm{x}$;

(*continued*)

Assignment 7 Outline Solution (continued 8)

A7 – 11. (cont.) (b) also, the estimates of γ_1 (the slope) and γ_0 (the intercept) of the (model) regression of **X** on **Y** are:

$$g_1 = 1.009\,476\,321, \qquad g_0 = -4.481\,153\,69, \qquad [1/g_1 = 0.990\,612\,636],$$

so the estimated regression of **X** on **Y** is: $\quad {}_{\text{reg}}\bar{x} = -4.48115 + 1.009\,48y,$

which is equivalent to: $\quad {}_{\text{reg}}\bar{y} = 4.439\,09 + 0.990\,613x.$

The two estimated regression lines are not the same because the estimated regression of **Y** on **X** is based on minimizing the squared *vertical* distances of the points from the line, whereas the estimated regression of **X** on **Y** is based on minimizing the squared *horizontal* distances.

(c) The coefficient of determination and the correlation coefficient are:

$$r^2 = 0.920\,956\,126, \qquad r = 0.959\,665.$$

- The value of the coefficient of determination shows that a little over 92% of the variation of the y_i's about their average has been accounted for by each estimated regression of **Y** on **X**;
- the value of the correlation coefficient, being *positive*, shows there is a positive association between x and y and, because the value is fairly close to its maximum possible value of 1, the points are fairly tightly clustered about each estimated regression line.

(d) The good fit of *both* the straight-line models to the data [as seen from visual inspection of the scatter diagram with the superimposed lines, and also as reflected by the high values of the coefficient of determination and the correlation coefficient in (c)] show that the two methods of dating are generally consistent with each other; however, we also see that there are appreciable discrepancies in a few individual cases – for instance, the four points marked on the scatter diagram by horizontal arrows, corresponding to the first, fifth, twentieth and last observations.

A7 – 12. (a) The scatter diagram of the U.K. radio receiver license data is shown below; for interest, the estimated regression of **Y** on **X** is also shown dashed on the diagram, based on the numerical quantities given below at the left.

$\Sigma x_i = 65.262,$

$\Sigma x_i^2 = 385.193\,966,$

$\Sigma y_i = 208,$

$\Sigma y_i^2 = 3{,}490,$

$\Sigma x_i y_i = 1148.08;$

$\bar{x} = 4.661\,571\,429,$

$\bar{y} = 14.\dot{8}\dot{5}\dot{7}\,\dot{1}\dot{4}\dot{2},$

$SS_{xy} = 178.473\,142,$

$SS_x = 80.970\,491\,4,$

$SS_y = 399.714\,285;$

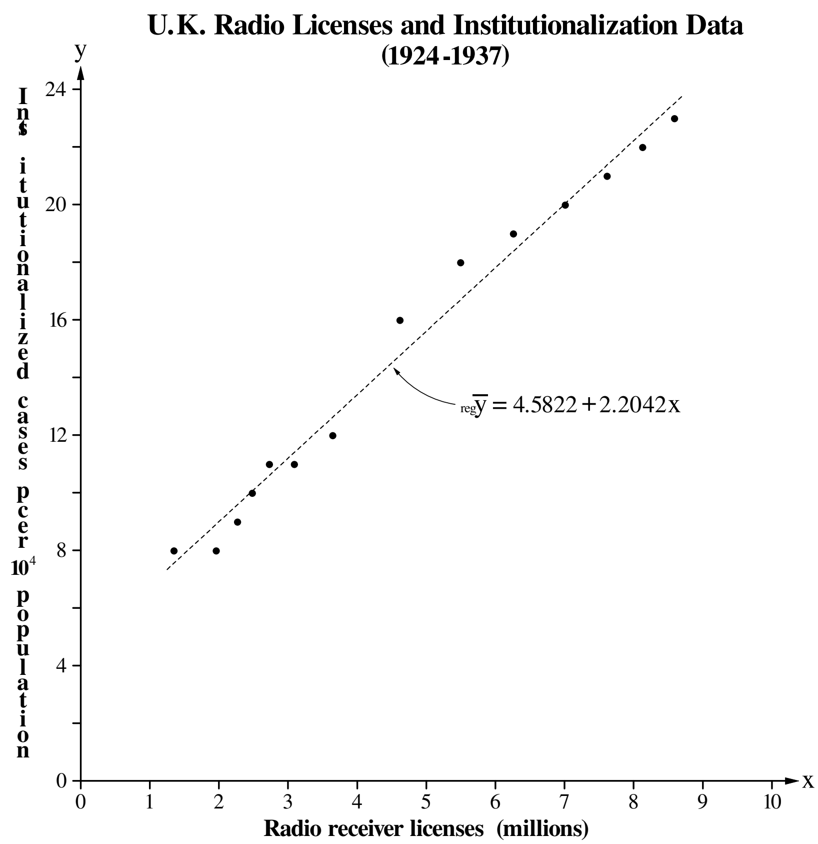
hence, the *estimates* of β_1 (the slope) and β_0 (the intercept) of the (model) regression of **Y** on **X** are:

$b_1 = 2.204\,175\,122,$

$b_0 = 4.582\,223\,083,$

so that the equation is:

${}_{\text{reg}}\bar{y} = 4.5822 + 2.2042x.$



(b) The correlation coefficient (and, for interest, the coefficient of determination) are:

$r = 0.992\,052\,483,$

$r^2 = 0.984\,168\,130.$

(c) The basis of the claim is the strong *association* the data show between the two variables – the points on the scatter diagram all lie reasonably close to the estimated regression line, and the value of the correlation coefficient in

(*continued overleaf*)

Assignment 7 Outline Solution (continued 9)

A7–12. (cont.) (c) (b) is positive and is close to its maximum possible value of 1. The *statistical* issue of importance is the implication of the claim that listening to the radio *causes* illness which may result in confinement to an institution; *i.e.*, the claim implies that the *reason* for the association is that one variable (here, listening to the radio) *causes* the other variable.

We know that association between two variables can arise for one of *three* reasons:

- a *causal connection* between the two variables;
- a third variable which is a *common cause* of the two variables;
- *coincidence*.

Because this investigation has an *observational* Plan, its data do *not* permit us to give the *first* reason rather than one of the other two, so that claim is *not* justified. In fact, it is more likely changes in the two variables are a common response to the numerous other societal changes in the U.K. over the period from 1924 to 1937.