University of Waterloo — STAT 221 – W. H. Cherry

Assignment 7

A7 – 1. As part of an investigation on the selection of grand juries in Alameda county, the educational level of grand jurors was compared with that of the county population as a whole, as shown at the right. Could equiprobable selecting of 62 people from the county show a distribution differing so widely from that of the county? Choose one of the following **five** options and justify your choice with an appropriate statistical test:

Educational level	Number of jurors	County
Elementary	1	28.4%
Secondary	10	48.5%
Some college	16	11.9%
College degree	35	11.2%
Total	**62**	**100.0%**

(a) this is absolutely impossible;
(b) this is possible but fantastically unlikely;
(c) this is possible but unlikely – the chance is around 1%;
(d) this is quite possible – the chance is around 10%;
(e) this is nearly certain.

A7 – 2. Someone claims to be rolling a pair of fair dice. To test their claim, you make them roll the dice 360 times and you observe the results given at the right. Would you accept an offer from this individual to play a game of chance with her or him using these dice? Set out your Answer in detail; it should include an appropriate test of fit.

Sum	2	3	4	5	6	7	8	9	10	11	12
Frequency	11	18	33	41	47	61	52	43	29	17	8

A7 – 3. The international Rice Research Institute in the Phillipines is developing new lines of rice which combine high yields with resistance to disease and insects. The technique involves crossing different lines to get a new line which has the most advantageous combination of genes: detailed genetic modelling is required. One project involved breeding new lines for resistance to an insect called *brown plant hopper*; 374 lines were raised, with the results shown below at the right. According to the IRRI genetic model, the lines are *independent*; each line has a 25% chance of being resistant, a 50% chance of being mixed, and a 25% chance of being susceptible. Are the results consistent with this model based on an assumption of **probabilistic independence**? Set out your Anser in detail.

	Number of lines
All plants resistant	97
Mixed: some resistant, some susceptible	184
All plants susceptible	93

A7 – 4. The distribution of *V1* flying bomb hits on the south of London during World War II was studied after the War. The region was divided into 576 squares of area ¼-square-kilometre, and the number of squares (n_k) receiving exactly k hits was as shown at the right.

k	0	1	2	3	4	5	6	7
n_k	229	211	93	35	7	0	0	1

(a) Test whether these data can reasonably be modelled by a Poisson distribution.
(b) Outline what can be concluded from the results of the test of fit in (a) about the view, popular in London during the War, that the points of impact of the flying bombs tended to cluster.

A7 – 5. In a pigmentation investigation of Scottish school children, the number of boys and girls whose hair colour fell into each of four classes were as shown at the right. Do these data provide adequate evidence that hair colour depends on sex? Set out your Answer in detail.

	HAIR COLOUR				
	Fair	*Red*	*Medium*	*Dark*	**Total**
Boys	60	12	85	53	210
Girls	54	10	68	38	170

A7 – 6. Four hundred students selected equiprobably were classified according to their smoking habits; they were then asked about the smoking habits of their parents. The results were as shown at the right.

STUDENT'S SMOKING HABITS	PARENTS' SMOKING HABITS			
	Neither smokes	*Only father smokes*	*Only mother smokes*	*Both smoke*
Non-smoker	115	70	29	18
Casual smoker	41	30	28	17
Regular smoker	20	11	10	11

(a) Do these data support the view that smoking habits of parents and their children are probabilistically independent? Set out your Answer in detail.
(b) Explain whether these data show that children tend to follow their parents' smoking habits.

A7 – 7. Text Exercise 9.17 (pages 649-650): *Alcohol and nicotine consumption during pregnancy may harm children. .....*

A7 – 8. Using seed from a single source, Gregor Mendel grew 529 pea plants and classified them according to seed shape (*round*, *round and wrinkled*, *wrinkled*) and seed colour (*yellow*, *yellow and green*, *green*). He obtained the following data:

38 round, yellow
60 round and wrinkled, yellow
28 wrinkled, yellow
65 round, yellow and green
138 round and wrinkled, yellow and green
68 wrinkled, yellow and green
35 round, green
67 round and wrinkled, green
30 wrinkled, green.

(*continued overleaf*)

A7 – 8. (a) Test the hypothesis the shape and colour classifications for the seeds are independent; set out your Answer in detail.
(b) According to Mendel's theory of heredity, the frequencies of yellow, yellow and green, and green seeds should be in the ratio 1:2:1. Test whether this hypothesis is supported by the data; set out your Answer in detail.

A7 – 9. (a) Text Exercise 2.99 (page 213): *The National Halothane Study was a major investigation of the safety.....*
(b) Text Exercise 2.107 (page 214): *Return to the investigation of the safety of anaesthetics from Exercise 2.99. .....*

A7 – 10. The following table gives the winning speeds (y mph) of the Indianapolis 500 auto race for the years 1919 through 1977 (x), except for the War years 1942-1945 when the race was not held:

Year	Speed	Year	Speed	Year	Speed	Year	Speed	Year	Speed	Year	Speed
1919	88.05	1929	97.56	1939	115.04	1949	121.33	1959	138.86	1969	156.87
1920	88.62	1930	100.45	1940	114.28	1950	124.00	1960	138.77	1970	155.75
1921	89.62	1931	96.62	1941	115.11	1951	126.24	1961	139.13	1971	157.74
1922	94.48	1932	104.11	1942	----	1952	128.92	1962	140.29	1972	162.96
1923	90.95	1933	104.16	1943	----	1953	128.74	1963	143.14	1973	159.04
1924	98.23	1934	104.86	1944	----	1954	130.84	1964	147.35	1974	158.59
1925	101.13	1935	106.24	1945	----	1955	128.21	1965	151.39	1975	149.21
1926	95.90	1936	109.07	1946	114.82	1956	128.49	1966	144.32	1976	148.73
1927	97.55	1937	113.58	1947	116.34	1957	135.90	1967	151.21	1977	161.33;
1928	99.48	1938	117.20	1948	119.81	1958	133.79	1968	152.88		

for these data: $\sum_{i=1}^{n} x_j = 107{,}158$, $\sum_{i=1}^{n} y_j = 6{,}837.28$, $\sum_{i=1}^{n} x_j y_j = 13{,}342{,}773.87$,
$\sum_{i=1}^{n} x_j^2 = 208{,}795{,}872$, $\sum_{i=1}^{n} y_j^2 = 877{,}972.23$, $n = 55$.

(a) Prepare a properly-labelled scatter diagram of these data.
(b) Find the estimated regression of $\mathbf{Y}$ on $\mathbf{X}$; show this line on your scatter diagram.
(c) Noting that the largest departures of the data from the fitted line tend to occur after the interruption of the race for the War years 1942-1945, suggest an extension of the *linear* model which would fit the data more satisfactorily.
(d) Use the estimated regression line to 'predict' the winning speed of the Indianapolis 500 in 1983 and in 2001; discuss briefly which of the two 'predictions' you would expect to be more accurate.

A7 – 11. Archeologists investigating Indian ruins in the southwestern United States have made extensive use of tree-ring dating as well as radiocarbon dating in estimating the age of artifacts. In one collection, the ages (in years) estimated by treering dating (x) and radiocarbon dating (y) were as follows:

x	710	717	350	323	500	620	832	669	917	423	212	822	612	647	513
y	795	764	320	360	612	642	786	690	878	436	222	765	543	642	533
x	722	724	400	396	812	415	272	204	206	824	641	527	569	693	471
y	724	745	409	456	652	432	352	187	192	764	701	529	582	646	360;

for these data: $\sum_{j=1}^{n} x_j = 16{,}743$, $\sum_{i=1}^{n} y_j = 16{,}719$, $\sum_{i=1}^{n} x_j y_j = 10{,}441{,}175$,
$\sum_{j=1}^{n} x_j^2 = 10{,}561{,}289$, $\sum_{i=1}^{n} y_j^2 = 10{,}417{,}377$, $n = 30$.

(a) Prepare a properly-labelled scatter diagram of these data.
(b) Find the estimated regressions of $\mathbf{Y}$ on $\mathbf{X}$ and $\mathbf{X}$ on $\mathbf{Y}$; show the two lines on your scatter diagram and explain briefly why they are different.
(c) Calculate the coefficient of determination and the correlation coefficient for the data; give their interpretations.
(d) Explain briefly whether the data indicate that the two methods of dating give *consistent* results.

A7 – 12. The following time series data, for the United Kingdom over the 14-year period 1924-37, show the number (in millions) of radio receiver licenses issued (x, say) and the number [per 10,000 estimated population (y, say)] of people confined to psychiatric institutions:

Year	1924	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937
x	1.350	1.960	2.270	2.483	2.730	3.091	3.647	4.620	5.497	6.260	7.012	7.618	8.131	8.593
y	8	8	9	10	11	11	12	16	18	19	20	21	22	23 ;

(*continued on the overleaf side of Assignment 8*)

Assignment 7 (continued)

A7 – 12. for these data: $\sum_{j=1}^{n} x_j = 65.262$, $\sum_{j=1}^{n} y_j = 208$, $\sum_{j=1}^{n} x_j y_j = 1{,}148.08$;

$\sum_{j=1}^{n} x_j^2 = 385.193\,966$; $\sum_{j=1}^{n} y_j^2 = 3{,}490$; $n = 14$.

(a) Prepare a properly-labelled scatter diagram of these data.

(b) Calculate the value of the correlation coefficient of x and y.

(c) Comment briefly on a claim based on the data that listening to the radio could result in confinement to a mental institution.