Lecture 2

August 1, 2023

1 Data

1.1 Data on USA Workforce composition

This data set comprises the figures published by the Labor Department for the composition of its workforce in 1986. It shows the average numbers over the year of male and female workers in the various different employment categories.

Type of employment	**Male (millions)**	**Female(millions)**
Professional	15.00	11.60
Industrial	12.90	4.50
Craftsman	12.30	1.25
Sales	6.90	6.45
Service	5.80	9.60
Clerical	3.90	14.30
Agricultural	2.80	0.60

Table 1: Average composition of the USA workforce during 1986

Some observations

1. Small data set
2. The number of males that are employed seemed to be more than the number of females.
3. Most number of people are employed in Professional category.
4. Least number of people work in Agriculture.

Questions

1. What is the most meaningful and appealing way to show the information?
2. How can we compare the male and female work forces in each category?
3. How to compare the total number of employees in each of the categories?

1.2 Data set on Infants with SIRDs

This Data set Comprises recorded birth weights of 50 infants who displayed severe idiopathic repiratorydistrss syndrome(SIRDs). This is a serious condition which can result in death.

1.050*	1.230*	1.500*	1.130	1.680	1.930
1.720*	1.770*	2.500*	2.090	2.700	3.160
1.100*	1.225*	1.295*	3.640	1.410	1.720
1.550*	1.890*	2.200*	2.200	2.550	1.262*
2.440*	2.730*	1.575	1.300*	1.940*	1.820*
1.760	2.015	2.600	2.270*	2.560*	2.275*
2.950	3.400	2.830	1.030*	1.185*	1.310*
1.715	2.040	2.400	1.600*	1.750*	1.175*
* child died					

Table 2: Birth weights (in kg) of infants with severe idiopathic respiratory distress syndrome

Some observations

1. The baby weights vary between 1.03 kg and 3.64 kg.
2. Children died due to this disease.

Questions

1. Can we relate the weight of the baby with the chances of survival?
2. Can we split the data into two identifiable groups?

1.3 Data set on Runners in The Tyneside Great North run.

This data set comes from 22 of the competitors in an annual championship run. The Tyneside Great North run. Blood Samples were taken from eleven runners before and after the run, and also from another eleven runners who collapsed near the end of the race. In table below, we see the measurements of beta endorphin concentrations in pmol/litre.

Normal runner before race	**Same runner after race**	**Collapsed runner after race**
4.3	29.6	66
4.6	25.1	72
5.2	15.5	79
5.2	29.6	84
6.5	24.1	102
7.2	27.8	110
8.4	20.4	123
9.0	21.9	144
10.4	14.2	162
14.0	34.6	169
17.8	40.2	414

Table 3: Blood plasma beta endorphin concentration

Some observations

1. The difference in the beta endorphin levels is quite visible before and after the race.
2. The difference between the beta endorphin levels of a runner who finished the race and the runner who collapsed is significantly high.

Questions

1. What is the relationship between beta endorphin concentrations before and after the race?
2. What is a typical concentration for a runner who finished the race?
3. What is typical concentration for a collapsed runner?
4. How do the dispersions of data values compare?

1.4 Data set on Alcoholism and Cirrhosis

This data shows the average alcohol consumption in litres per person per year and the death rate per 100 000 of the population from cirrhosis and alcoholism.

Country and Alcohol	**Consumption (1/person/year)**	**Cirrhosis and Alcoholism (death rate/100,000)**
France	24.7	46.1
Italy	15.2	23.6
W. Germany	12.3	23.7
Austria	10.9	7.0
Belgium	10.8	12.3
USA	9.9	14.2
Canada	8.3	7.4
England	7.2	3.0
Sweden	6.6	7.2
Japan	5.8	10.6
Neatherlands	5.9	3.7
Ireland	5.6	3.4
Norway	4.2	4.3
Finland	3.9	3.6
Israel	3.1	5.4

Table 4: Average alcohol consumption and death rate

Some observations

1. The two variable seem to be related to each other.
2. France has a noticeably higher average annual individual alcohol consumption than the others.
3. The French alcohol-related death rate is highest as well.

Questions

1. What is the relationship between alcoholism and death rate and is it strong one?
2. How can the strength of such a relationship be measured?

1.5 Data set on Body and brain weights for animals

This data set comprises average body and brain weights for 28 kinds of animal, some of them extinct.

Species	Body weight (kg)	Cirrhosis Brain (g)
Mountain Beaver	1.350	8.100
Cow	465.000	423.000
Grey Wolf	36.330	119.500
Goat	27.660	115.000
Guinea Pig	1.040	5.500
Diplodocus	11700.000	50.000
Asian Elephant	2547.000	4603.000
Donkey	187.100	419.000
Horse	521.000	655.000
Potar Monkey	10.000	115.000
Cat	3.300	25.600
Giraffe	529.000	680.000
Gorilla	207.000	406.000
Human	62.000	1320.000
African Elephant	6654.000	5712.000
Triceratops	9400.000	70.000
Rhesus Monkey	6.800	179.000
Kangaroo	35.000	56.000
Hamster	0.120	1.000
Mouse	0.023	0.400
Rabbit	2.500	12.100
Sheep	55.500	175.000
Jaguar	100.000	157.000
Chimpanzee	52.160	440.000
Brachiosaurus	87000.000	154.500
Rat	0.280	1.900
Pig	192.000	180.000

Table 5: Average body and brain weights for animals

Some observations

1. The two seem to be related.

Questions

1. What is the relationship between body weight and brain weight and how strong is it?

2. Can the strength of the relationship be measured?

3. Is a larger brain really required to govern a larger body?

1.6 Data set on Stock-market averages

This is the list of the annual highs and lows for the Dow Jones industrial average on the New York stock-market from 1954 to 1985.

Table 6: Dow Jones industrial averages

Year	**High**	**low**
1954	404	280
1955	488	388
1956	521	462
1957	521	420
1958	584	437
1959	679	574
1960	685	566
1961	735	419
1962	726	536
1963	767	647
1964	892	766
1965	969	841
1966	995	744
1967	943	786
1968	985	825
1969	969	770
1970	842	631
1971	951	798
1972	1036	889
1973	1052	0.788
1974	892	578
1975	882	632
1976	1015	859
1977	1000	801
1978	908	742
1979	898	797
1980	1000	759
1981	1024	824
1982	1071	777
1983	1287	1027
1984	1287	1087
1985	1360	1185

Some observations

1. Starting with 1954, you see a steady increase followed by a decrease which does not go as low as the 1954 value before it starts to increase again. After a little fluctuation, it increases to the 1968 value, drops again, hits a high in 1972, 1976, and so on.

2. The values show an overall increase over the 32-year period.

Questions

1. Can we use these data to predict highs and lows of the Dow Jones average for 1986 and subsequent years?

1.7 Data set on Surgical removal of tattoos

This data set comprises of clinical data from 55 patients who have had forearm tattoos removed by two different surgical methods. Their tattoos were of large, medium or small size, either deep or at moderate depth. The final result is scored from 1 to 4, where 1 represents a poor removal and 4 represents an excellent result. The two methods of removal are denoted A and B. The sex of the patient is also shown.

Method	Sex	Size	Depth	Score
A	M	large	deep	1
A	M	large	moderate	2
B	F	small	deep	1
B	M	small	moderate	1
B	F	large	deep	4
B	M	medium	moderate	3
B	M	medium	deep	4
A	M	large	deep	4
A	M	large	moderate	1
A	M	small	moderate	4
A	M	large	deep	4
A	M	large	moderate	3
A	F	small	moderate	3
B	M	large	deep	2
B	M	large	deep	2
B	F	medium	moderate	1
B	M	large	deep	1
B	F	medium	deep	
B	F	small	moderate	3
A	F	small	moderate	4
B	M	large	deep	2
A	M	medium	moderate	4
B	M	large	deep	4
B	M	large	moderate	4

Table 7: Surgical removal of tattoos

Some observations

1. The data are not numeric.

Questions

1. What are the relative merits of the two methods of tattoo removal?
2. Is one method simply better, or does it depend upon the size or depth of the tattoo?

2 Reference

Daly, F., Hand, D.J., Jones, M.C., Lunn, A.D. and McConway, K.J., 1995. Elements of statistics. Addison-Wesley Publishing Company.