11/12/18

Basic Data Science (15 No. Questions)

1)

No. of children	No. of families
0	8
1	16
2	22
3	14
4	6
5	4
6	2

(a) Calc. the mean no. of children/family for the sample from the given table.

(b) Calc. S.D

Sol) (a)

No. of children (x_i)	No. of families (f_i)	$f_i \cdot x_i$	$f_i \cdot x^2$
0	8	0	0
1	16	16	16
2	22	44	88
3	14	42	126
4	6	24	96
5	4	20	100
6	2	12	72
	$\sum f_i = 72$	$\sum f_i x_i = 158$	$\sum f_i \cdot x_i^2 = 498$

By using freq. distribution, we get all.

$\therefore$ Mean $\mu = \frac{\sum f_i x_i}{\sum f_i} = \frac{158}{72} = 2.1944 \approx 2.19$.

(b) Variance (Population), $\boxed{\sigma^2 = \frac{\sum (x_i - \mu)^2 \cdot f_i}{\sum f_i}}$ for Grouped data

$$\sigma^2 = \frac{(0-2.19)^2 \cdot 8 + (1-2.19)^2 \cdot 16 + (2-2.19)^2 \cdot 22 + (3-2.19)^2 \cdot 14 + (4-2.19)^2 \cdot 6 + (5-2.19)^2 \cdot 4 + (6-2.19)^2 \cdot 2}{72}$$

$$\Rightarrow \sigma^2 = \frac{38.3688 + 22.6576 + 0.7942 + 11.34 + 13.14 + 40.7044 + 35.1122}{72}$$

$$\Rightarrow \sigma^2 = \frac{162.1172}{72} = 2.2516$$

S.D, $\sigma = \sqrt{\sigma^2} = \sqrt{2.2516} = 1.5$.
(Population)

2) A national random sample of 20 ACT scores from 2010 is listed below.

29, 26, 13, 23, 23, 25, 17, 22, 17, 19, 12, 26, 30, 30, 18, 14, 12, 26, 17, 18.

(a) Calc. the sample mean & S.D

(b) Calc no. of observations that are 2 or more sample S.D from the sample mean.

(c) Provided that the ACT is reasonably normally distributed with a mean of 18 & S.D of 6, determine the proportion of students with 33 or higher.

Sol) (a) Sample Mean, $\bar{x} = \frac{\sum x}{n}$, n is size of sample

$$\bar{x} = \frac{417}{20} = 20.85.$$

Sample S.D, $s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$

$$\Rightarrow s = \sqrt{\frac{(29-20.85)^2 + (26-20.85)^2 + (13-20.85)^2 + (23-20.85)^2 + (23-20.85)^2 + (25-20.85)^2 + (17-20.85)^2 + (22-20.85)^2 + (17-20.85)^2 + (19-20.85)^2 + (12-20.85)^2 + (26-20.85)^2 + (30-20.85)^2 + (30-20.85)^2 + (18-20.85)^2 + (14-20.85)^2 + (12-20.85)^2 + (26-20.85)^2 + (17-20.85)^2 + (18-20.85)^2}{20-1}}$$

$$\Rightarrow s = 5.94.$$

(b) We have to identifying any data points that fall outside the range of $\bar{x} \pm 2s$ (mean $\pm$ 2 x S.D).

So, the boundaries are -

$$20.85 \pm (2 \times 5.94) \Rightarrow 20.85 \pm 11.88.$$

i.e., $20.85 - 11.88$ and $20.85 + 11.88$

i.e., 8.97 and 32.73

On checking each data point, none of them are falling outside this range.

No observation is beyond this range, so, answer is 0.

(c) Proportion of students with a score of 33 or higher in a normal dist^{n}:

Here, $\mu = 18$, $\sigma = 6$.

Let X be the random variable denoting ACT scores, which is normally distributed,

Using z-score formula,

$$z = \frac{x - \mu}{\sigma} = \frac{x - 18}{6}$$

Probability of students with a score of 33 or higher:

$P(X \geq 33)$

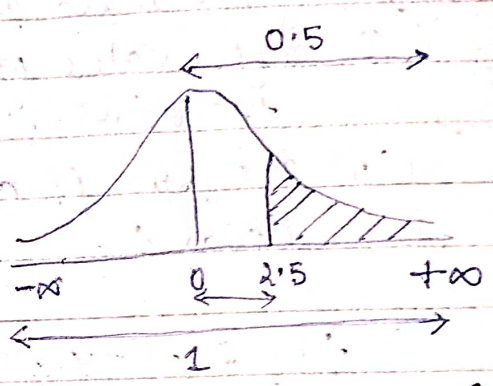
$\Rightarrow P(z \geq 33)$ $\Rightarrow P\left(z \geq \frac{33-18}{6}\right)$

$\Rightarrow P(z \geq 2.5)$

$\Rightarrow 0.5 - P(0 \leq z \leq 2.5)$

$\Rightarrow 0.5 - 0.49379$

$\Rightarrow 0.00621$



∴ So, the proportion of students with a score of 33 or higher is approximately $0.0062 \approx 0.62\%$

{ ∵ From z-table(tve)
$P(z \leq 2.5) = 0.9938$
$\Rightarrow 0.9938 - 0.5$
$P(0 \leq z \leq 2.5) = 0.4938$.

3) (b) The no. of ships to arrive at a harbor on any given day is a random var. represented by x. The P.D for x is:

x	10	11	12	13	14
$P(x)$	0.4	0.2	0.2	0.1	0.1

Find the prob. that on a given day:

(i) exactly 14 ships arrive

(ii) Atleast 12 ships arrive

Sol) (i) $P(x = 14) = 0.1$

So, the probability that on a given day exactly 14 ships arrive is 0.1.

(ii) At least 12 ships arrive i.e.,

$P(X \geq 12) = P(X=12) + P(X=13) + P(X=14)$

$= 0.2 + 0.1 + 0.1$

$= 0.4$.

$\therefore$ Probability $= 0.4$.

4) (b) Which of the following are prob. functions,

(i) $f(x) = 0.25$ for $x = 9, 10, 11, 12$.

Sol) For a function to be valid probability function, it must satisfy 2 conditions: (Both)

(1) Each value of the function must be between 0 and 1, inclusive.

(2) The sum of all possible values of the function must equal 1.

For $f(x) = 0.25$, for $x = 9, 10, 11, 12$

$x =$	9	10	11	12
$f(x) =$	0.25	0.25	0.25	0.25

Here conditions 1 is satisfy., i.e., every value is b/w $0 \leq f(x) \leq 1$

Now, The sum of all possible values is

$0.25 + 0.25 + 0.25 + 0.25 = 1$, so condition 2 satisfies.

$\therefore$ (i) is a valid Probability function.

(ii) $f(x) = \frac{(3-x)}{2}$, for $x = 1, 2, 3, 4$.

x	1	2	3	4
$f(x)$	1	0.5	0	-0.5

Here, condition 1 is not satisfied, i.e., for $f(x)$; $x = 4$ the value is -0.5 which is not in range of $0 \leq f(x) \leq 1$.

$\therefore$ (ii) is not a valid Probability function,

(iii) $f(x) = \frac{(x^2+x+1)}{25}$, for $x = 0, 1, 2, 3$

x	0	1	2	3
f(x)	1/25	3/25	7/25	13/25

Here, condition 1 is ~~not~~ satisfied

And, $\sum f(x) = \frac{1}{25} + \frac{3}{25} + \frac{7}{25} + \frac{13}{25}$

$= \frac{24}{25} \approx 1$., so condition 2 also satisfied.

(5) A survey asked people how often they exceed speed limits. The data are then categorized into the following contingency table of counts showing the relationship b/w age group & response

	Exceed Limit if possible?		
Age	Always	Not Always	Total
Under 30	100	100	200
Over 30	40	160	200
Total	140	260	400

(a) Among people with age over 30, what's the "risk" of always exceeding the speed limit?

(b) Among people with age under 30, what are the odds that they always exceed the speed limit?

(c) What is the relative risk of always exceeding the speed limit for people under 30 compared to people over 30?

Sol) Using Contingency table,

(a) Risk $= \frac{\text{No. of People Over 30 Always Exceeding}}{\text{Total no. of People Over 30}}$

$= \frac{40}{200} = 0.20$

So, the risk of always exceeding the speed limit for people over 30 is 0.20.

(b) Odds of always Exceeding Speed limit for people under 30:

$$\text{Odds} = \frac{\text{No. of People Under 30 Always Exceeding}}{\text{No. of People Under 30 Not Always Exceeding}}$$

$$\text{Odds} = \frac{100}{100} = 1 \text{ to } 1.$$

~~So, the~~

(c) Relative Risk of Always Exceeding Speed limit for People Under 30 compared to people Over 30:

$$\text{Relative Risk} = \frac{\text{Risk for People Under 30}}{\text{Risk for People Over 30}}$$

$$= \frac{\frac{100}{200}}{\frac{40}{200}} = \frac{\cancel{100}^{5}}{\cancel{200}} \times \frac{\cancel{200}}{\cancel{40}_{2}} = 2.5.$$

7) (a) Following sample of no. of coffee sales per day in a cafe:

550, 570, 490, 615, 505, 580, 570, 460, 580, 530, 526

If t-test is used assuming normal pop^{n} then calc. value of t.

Sol) For calculating t-test's 't', the formula is (for one-sample test)

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$, where $\bar{x}$ → sample mean
μ → Population "
s → Sample S.D
n → sample size

Let us assume population mean (μ_0) = 0.

i.e., $H_0: \mu_0 = 0$

& $H_a: \mu_0 \neq 0$.

As, $n = 11$, $\mu_0 = 0$ and population S.D is unknown

So, we are using t-test.

x	$(x-\bar{x})$	$(x-\bar{x})^2$
550	6.73	45.29
570	26.73	714.49
490	-53.27	2837.69
615	71.73	5195.19
505	-38.27	1464.59
580	36.73	1349.09
570	26.73	714.49
460	-83.27	6933.89
580	36.73	1349.09
530	-13.27	1760.09
526	-17.27	298.25
$\sum x = 5976$		22611.97

$\therefore$ sample mean, $\bar{x} = \frac{\sum x}{n} = \frac{5976}{11} = 543.27$

sample S.D, $s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$

$= \sqrt{\frac{22611.97}{10}} = \sqrt{2261.197} = 47.55$

On substituting all,

$t = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{543.27-0}{\frac{47.55}{\sqrt{11}}} = \frac{543.27}{14.33} = 37.91$

9) 29, 26, 13, 23, 23, 25, 17, 22, 17, 19, 12, 26, 30, 30, 18, 14, 12, 26, 17, 18.
Calc. the 95% confidence interval for the mean ACT score based on the t-distribution.
If mean (x) is 4 & the dist^n is 2, 3, 4, 5, 6, then calc. sum of squared deviations from the mean.

Sol) (a) Mean $\bar{x} = \frac{\sum x}{n} = \frac{417}{20} = 20.85$.
(Population / sample)

Population S.D, $\sigma = \sqrt{\frac{\sum (x-\bar{x})^2}{N}}$

$$\sigma^2 = \frac{(29-20.85)^2+(26-20.85)^2+(13-20.85)^2+(23-20.85)^2+(23-20.85)^2}{+(25-20.85)^2+(17-20.85)^2+(22-20.85)^2+(17-20.85)^2+(19-20.85)^2+}$$
$$\frac{(12-20.85)^2+(26-20.85)^2+(30-20.85)^2+(30-20.85)^2+(17-20.85)^2+(18-20.85)^2}{+(14-20.85)^2+(12-20.85)^2+(26-20.85)^2+(17-20.85)^2+(18-20.85)^2}$$
$$20$$

$$\sigma^2 = \frac{(9.85)^2+(6.85)^2+(-7.85)^2+(3.85)^2+(3.85)^2+(5.85)^2+(-3.85)^2}{+(2.85)^2+(-3.85)^2+(-1.85)^2+(-8.85)^2+(6.85)^2+(10.85)^2}$$
$$\frac{+(10.85)^2+(-2.85)^2+(-6.85)^2+(-8.85)^2+(6.85)^2+}{(-3.85)^2+(-2.85)^2}$$
$$20$$

$$\sigma^2 = \frac{874.55}{20} = 43.7275$$

$$\therefore \text{S.D}, \sigma = \sqrt{\sigma^2} = \sqrt{43.7275} = 6.612$$

As we know; Taking confidence Interval level 95% i.e., $z = 1.960$

$$\because \alpha = 1 - \frac{C}{100} = 1 - \frac{95}{100} = 0.05.$$

$$\Rightarrow z_{\alpha/2} = z_{0.05/2} \Rightarrow z_{0.025} = 1.96. \quad [\text{from z-table}]$$

$$\therefore \text{Confidence Interval, C.I} = \bar{x} \pm \left[z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right]$$

$$= 20.85 \pm \left[1.96 \times \frac{6.612}{\sqrt{20}}\right]$$

$$= 20.85 \pm [1.96 \times 1.47]$$

$$= 20.85 \pm 2.89.$$

6.27
2.78.

i.e., $[20.85 - 2.89, 20.85 + 2.89]$

6.78

C.I is $[17.96, 23.74]$.

(b) Given a set of numbers: 2, 3, 4, 5, 6 with a $\bar{x}$ of 4:

Sum of Squared Deviations from the Mean =

$$= (2-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(6-4)^2$$
$$= (-2)^2+(-1)^2+0+(1)^2+(2)^2$$
$$= 4+1+1+4 = 10$$

12) (1, 1·24), (2, 5·23), (3, 7·24), (4, 7·60), (5, 9·97), (6, 14·31), (7, 13·99), (8, 14·88), (9, 18·04), (10, 20·70).

(a) Given the following data pairs (x, y), find the regression equation.

(b) Calc. the correlation coefficient for the dataset

(c) For the following dataset, obtain a prediction for x = 4·5.

Sol) (a) To find the regression eqⁿ for the given data pairs (x, y), the slope (m) & y-intercept (b) in eqⁿ $y = mx + b$.

$$\text{Slope, } m = \frac{n(\sum x \cdot y) - (\sum x)\cdot(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Firstly, n = 10 (total size)

$\sum x = 1+2+3+4+5+6+7+8+9+10 = 55$

$\sum y = 1·24+5·23+7·24+7·60+9·97+14·31+13·99+14·88+18·04+20·70 = 113·2$

$\sum x \cdot y = (1\times1·24)+(2\times5·23)+(3\times7·24)+(4\times7·60)+(5\times9·97)+(6\times14·31)+(7\times13·99)+(8\times14·88)+(9\times18·04)+(10\times20·70) = 785·86$

$\sum x^2 = 1+4+9+16+25+36+49+64+81+100 = 385$

On substituting all values,

*) $\text{Slope, } m = \frac{10\times785·86 - (55\times113·2)}{(10\times385) - (55)^2}$

$= \frac{7858·6 - 6226}{3850 - 3025} = \frac{1632·6}{825} = 1·98$

*) $\text{Y-Intercept } (b) = \frac{\sum y - m(\sum x)}{n}$

$= \frac{113·2 - 1·98\times55}{10} = \frac{4·3}{10} = 0·43$

∴ On substituting all values on eqⁿ $y = mx + b$ we get, $y = 1·98x + 0·43$.

(b) Correlation coefficient, $r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$

So, $\sum y^2 = (1·24)^2 + (5·23)^2 + \dots + (20·70)^2$

$= 1614·3$

$$r = \frac{10 \times 785.86 - (55 \times 113.2)}{\sqrt{[(10 \times 385) - 3025] \times [10 \times 1614.3 - 12814.24]}}$$

$$= \frac{1632.6}{\sqrt{825 \times 3328.76}} = \frac{1632.6}{\sqrt{2746227}} = 0.985$$

(c) On putting $x = 4.5$ on eqⁿ $y = 1.98x + 0.43$ which we got in ques (a).

$\Rightarrow y = 1.98 \times 4.5 + 0.43$

$\Rightarrow y = 9.34$

15

Blood test	Bowel Cancer		→ Actual data
	Yes	No	
+ve	2 (TP)	(FP) 18	
-ve	1 (FN)	(TN) 132	

↑ Predicted data

(a) Calc. the sensitivity

Sol) Sensitivity (True Positive/TP or Recall), it is used to calculate the models ability to predict positive values. Here, the proportion of actual +ve cases correctly identified by a diagnostic test.

$$\text{Sensitivity} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$$

$$= \frac{2}{2+1}$$

$$= \frac{2}{3} = 0.6667 \approx 66.67\%$$

{ TP = 2, FN = 1

(b) False positive rate (False Alarm Rate or Type I Error Rate) is the proportion of actual negative cases that are incorrectly identified as +ve by a diagnostic test

$$\text{False +ve rate} = \frac{FP}{FP + TN}$$

$$= \frac{18}{18 + 132} = \frac{18}{150} = 0.12$$

$\approx 12\%$

{ FP = 18, TN = 132

(c) Using Bayes Theorem,

$$P(\text{cancer}|\text{Positive}) = \frac{P(\text{Positive}|\text{cancer}) \times P(\text{cancer})}{P(\text{Positive})}$$

where, $P(\text{Positive}|\text{cancer}) = \frac{TP}{TP+FN} = 0.66$

$$P(\text{cancer}) = \frac{TP+FN}{\text{Total}} = \frac{2+1}{153} = 0.019$$

$$P(\text{Positive}) = \frac{TP+FP}{\text{Total}} = \frac{2+18}{153} = 0.13$$

On substituting all,

$$\Rightarrow P(\text{cancer}|\text{Positive}) = \frac{0.66 \times 0.019}{0.13} = \frac{0.012}{0.13} = 0.096 \approx 0.1$$

(18) Compute Information Gain.

Outlook	+class	-class
Sunny	2	3
Overcast	4	0
Rainy	3	2

Also calc. entropy for the attribute Outlook

Sol) For attribute Outlook, the entropy,

first, we have +class = 9, -ve = 5, total = 14

$$\therefore \text{Entropy} = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$= 0.4097 + 0.5305 = 0.9402$$

For Sunny, the entropy,

We have, +ve class = 2, -ve = 3, total = 5

$$\therefore \text{Entropy} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$= 0.5288 + 0.4422 = 0.971$$

For Overcast, the entropy, here, +ve = 4, -ve = 0, total = 4

$$\therefore \text{Entropy} = -\frac{4}{4}\log_2\frac{4}{4} = 0.$$

For Rainy, the entropy, given +ve = 3, -ve = 2, total = 5

$$\therefore \text{Entropy} = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.5288 + 0.4422$$

$$= 0.971$$

Average weighted value $= \sum \frac{p_i + n_i}{p+n}$

$= \frac{2+3}{14} \times 0.971 + \frac{4+0}{14} \times 0 + \frac{3+2}{14} \times 0.971$

$= 0.6935 \approx 0.694$

∴ Information Gain for Outlook = Entropy of entire set − Avg. weighted value

$= 0.940 - 0.694 = 0.246$