LEC 15

Decimal Scaling

$$\text{Scaled value} = \frac{\text{Original V}}{10^j}$$

j = smallest int. $\max(\text{scaled value}) < 1$

$10^j \rightarrow$ Scaling factor.

1- Find max absolute value
2- Find j such that 10^j brings maximum absolute value below 1
3- Scale data.

Value: 1200, 5600, 8900, 3400, 10,000

1- max value: 10,000

2- $\frac{10,000}{10^j} < 1 \Rightarrow$ j: 4 $\quad \frac{10,000}{10^4} = 1$ X No

$j = 5 \quad \frac{10,000}{10^5} = 0.1$

So $j = 5$.

Lets scale

1200	0.012
5600	0.056
8900	0.089
3400	0.034
10000	0.1

When data set contain larger value → Use this

Sensitive to outliers as j may become very large

Smoothing

Replacing noisy data.

Day	Temp
1	22.5
2	23.1
3	24.0
4	100.0
5	23.8
6	24.2
⋮	
10	24.8

49.3

outlier. Lets smoothi noise.

e.g we use zscore or IQR to detect outlier.

Lets do Binning. Bins = 3 (Bin size) 3 days.

10/3 = 3 or 4.

Bins	Days	Temp	Mean
1	1,2,3	22.5, 23.1, 24.0	23.2
2	4,5,6	100, 23.8, 24.2	49.3
3	7,8,9	23.5, 24.5, 25.0	24.8
4	10	24.8	24.8

Normalization. $\frac{95-80}{95-80}$

1- Min max → When have diff want to bring to comon

$X = 2, 5, 8, 12 + \cdots$

$X' = 0, 0.06, 0.13, 0.21$

Normalized data is easier to u

2- Z score.

$$\mu = \frac{2+5+8+12+\cdots}{10} =$$

$$\sigma = \sqrt{\frac{(x-\mu)^2}{n}}$$

Z score →

2	→	-1.3
5		-1.1
8		-0.9
12		-0.6
18		-0.3
⋮		⋮

80 30 $\frac{95-80}{30}$ = 1

Normalization.

$\frac{95-80}{95-80}$

1- Min max. → When have diff scales in features. you want to bring to comon range. KNN, NN, SVM

X = 2, 5, 8, 12 +.... $X' = \frac{X-2}{50-2}$

X' = 0, 0.06, 0.13, 0.21

Normalized data is easier to visulize.

2- Z score.

$\mu = \frac{2+5+8+12+\dots}{} = 23.3$

$\frac{60-30}{2}$ 215 → B1 = 50–65 L.Income
65–80 High

1- Bin means

B1 = M = $\frac{2+5+8}{3}$ = 5 → [5, 5, 5]

B2 = [21.25, 21.25, 21.25, 21.25]

B3 = [44.33, 44.33, 44.33]

2- Bin boundaries:

B 1 = [2, 5, 8] → [2, 2, 8]

[12, 13, 30, 30]

[38, 38, 50

OR

Income.

1st	50–60	Avg Low
2nd	70–800	High

Bin range	Values
2 - 18	2, 5, 8, 12, 18
19 - 34	25, 30
35 - 50	38, 45, 50

- Equi depth (Based on sorted data)

Total value = 10

Bin size = 10/3

Bin	values
1st	2, 5, 8
2nd	12, 18, 25, 30
3rd	38, 45, 50

e.g Grouping customers purcha
amount into bins where ea
bin contain equal
no of
customers

Smoothing

1- Bin means

B1 = M =

sensitive to outliers.

Range = 50 - 2 = 48

Bin width = 48/3 = 16

Bin range	values
2 - 18	2, 5, 8, 12,
19 - 34	25, 30
35 - 50	38, 45, 5

- Equi depth (on based on desired d

Total value =

Bin size = 1

Bin	values

Lets take dataset

2, 5, 8, 12, 18, 25, 30, 38, 45, 50

1- Equi width (on normal data)

No of bins = 3

↓

Square root method.

$$K = \sqrt{n}$$

$$K = \sqrt{10} = 3.16 \simeq 3$$

$7.396 < 7.81$ we fail to reject H_0 at $\rho = 0.05$.

We cant conclude that drink affects energy levels.

Stats based on

For Whole dataset

Water: Energ

$$\frac{(5-7.25)^2}{7.25} = \frac{5.06}{7.25} = 0.698$$

Water: Tired

$$\frac{(5-2.75)^2}{2.75} = 1.854$$

Coffee - E

$$\frac{(8-7.25)^2}{7.25} = 0.077$$

C - T = 0.204

Tea E = 0.215 ; T - T 0.567

E-D E = 1.045 ; E-D-T = 2.75

5-1 $X^2 = 0.698 + 1.84 + 0.077 + 0.204 + 0.215 + 0.567 +$

$$\boxed{X^2 = 7.396}$$

5- Compare with Chi table.

df	P=0.10	P=0.05 (sig)	P=0.01 (U.sig)
3	6.25	7.81	11.34

d) = (4-1) + (2-1) = 3

Lets calculate 3 expected value manually.

$$E = \frac{R.T \times C.T}{G.T} =$$

3- Energetic.

1) - $E = \frac{10 \times 29}{40} = \frac{290}{40} = 7.25$

2) $E = \frac{10 \times 29}{40} = 7.25$

3) 7.25

Now rest will be determined automatically.

4- Expected.

	Energ.	Fluct.	total.
N	7.25	10 - 7.25 = 2.75	10
C	7.25	2.75	10
T	7.25	2.75	10
E.D	7.25	2.75	10
Total.	29.	11	40

5- Chi-Square.

$$X^2 = \sum \frac{(O-E)^2}{E}.$$

Cell by cell calculation.

3 df Example.

A researcher wants to test whether the type of drink effects energy level.
They conduct experiment on 40 people.
and divides them in 4 groups based on drink they consumed.

Water, coffee, Tea, Energy drink.

$H_0 \rightarrow$ type of drink has no effect.
all due to random chance.

$H_1 \rightarrow$ type of drink consumed effect a person energy.

2) Actual table.

	Energetic	Tired	Total.
W	5	5	10
C	8	2	10
Tea	6	4	10
E.D	10	0	10
Total.	29	11	40

0.05, meaning we CANNOT say for sure that carrots help running speed.

The Grand Professor nodded.

> "Yes, Charlie! This means we do not have enough evidence to say carrots actually help. The difference could have happened just by chance."

Charlie now understood:

- ☑ df controls which Chi-Square Table we use.
- ☑ p-value tells us if the result is significant.
- ☑ If p < 0.05, we reject H_0 (the null hypothesis) and say it's significant.
- ☑ If p > 0.05, we fail to reject H_0, meaning there's not enough proof.

Calculate Cramer's V strength of association b/w variables

$V = \sqrt{\frac{\chi^2}{n \times (k-1)}} = \sqrt{\frac{3.84}{20 \times (2-1)}} = 0.48$

Epilogue: The Final Lesson

Charlie ran back to Mathville and told everyone:

> "Statistics is like detective work! You can't just guess—you have to prove if something is real or just random!"

And from that day on, Mathville became the most logical place in the world.

The End.

Key Takeaways:

Term	Meaning
p-Value	Probability that your result happened just by chance
Significant Value (Critical Value)	The cutoff number that decides if your result is "real"
p < 0.05	We call the result statistically significant ☑
p > 0.05	The result is not significant—it could be just luck ✖

$$df = (\text{Rows} - 1) \times (\text{Columns} - 1)$$

Here, we have 2 rows and 2 columns, so:

$$df = (2 - 1) \times (2 - 1) = 1$$

Charlie checked df = 1 in the Chi-Square Table:

df	p = 0.10	p = 0.05 (Significant)	p = 0.01 (Very Significant)
1	2.71	3.84	6.63

His chi-square value was 3.34, which was less than 3.84.

📌 The p-value is greater than 0.05!

Final Level: What Does This Mean?

Charlie had solved the case! 🕵

> "Because our chi-square number (3.34) is less than 3.84, the p-value is greater than 0.05, meaning we CANNOT say for sure that carrots help running speed!"

The Grand Professor nodded.

> "Yes, Charlie! This means we do not have enough evidence to say carrots actually help. The difference could have happened just by chance."

Charlie now understood:

- ☑ df controls which Chi-Square Table we use.
- ☑ p-value tells us if the result is significant.
- ☑ If p < 0.05, we reject H_0 (the null hypothesis) and say it's significant.
- ☑ If p > 0.05, we fail to reject H_0, meaning there's not enough proof.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Where O is observed and E is expected."

Charlie calculated:

$$\chi^2 = \frac{(8-6)^2}{6} + \frac{(2-4)^2}{4} + \frac{(4-6)^2}{6} + \frac{(6-4)^2}{4}$$

$$= \frac{4}{6} + \frac{4}{4} + \frac{4}{6} + \frac{4}{4} = 0.67 + 1 + 0.67 + 1 = 3.34$$

"Great job, Charlie! Now, to get the p-value, we need the Chi-Square Table!"

Level 3: The p-Value and the Significance Test 🔥

Charlie ran to the Chi-Square Table and found his degrees of freedom (df).

"Professor, how do I get df?" 😕

The professor explained,

"You count how much your data can change before one value becomes fixed. Use this formula:

$$df = (\text{Rows} - 1) \times (\text{Columns} - 1)$$

Here, we have 2 rows and 2 columns, so:

$$df = (2-1) \times (2-1) = 1$$

Charlie checked df = 1 in the Chi-Square Table:

Level 3: The Count of Degrees of Freedom (df)

The Game of Numbers: The Secret of Degrees of Freedom and p-Value

Once upon a time, in the land of **Mathville**, there was a fun game played by scientists and detectives. This game helped them find out if something was real or just random luck.

Our hero, **Charlie Square**, loved solving number mysteries. One day, the **Grand Professor of Numbers** gave him a new case:

> **"Charlie, people are claiming that eating carrots helps kids run faster! But we don't know if this is actually true or just a random coincidence. Your mission: Find the truth!"**

Charlie was excited! But how could he prove whether carrots really made kids faster?

H_0: Eating carrots does not improve running speed.
H_1 → opposite

Level 1: The Running Race

Charlie gathered **20 kids** and made them **run a race**.

- **10 kids ate carrots** before running.
- **10 kids did not eat carrots** before running.

Charlie wrote down their times and wanted to see if **carrot-eaters really ran faster**. But how?

The **Grand Professor** handed Charlie a **magic tool** called the **Chi-Square Test**!

> **"Charlie, you can't just look at the times and guess. You need to test if the difference is real or just random chance!"**

Charlie was ready!

Level 2: The Secret of Degrees of Freedom (df)

Step deviation

Step h:10
Since values have constant gap of 10.

$$dx = \frac{X - 30}{h}$$

$$dy = \frac{Y - (33)}{h}$$

assumed

Uniform gap there more so we use simple.

dx	dy	dx.dy	dx^2	dy^2
-2	-2.1	4.2	4	4.41
-1	-0.8	0.8	1	0.64
0	0	0	0	0
1	1.2	1.2	1	1.44
2	2.2	4.4	4	4.84

$$r = \frac{10.6}{\sqrt{10 \times 11.33}}$$

$$= 0.997$$

Useful in grouped data or when nos are larger.

Assumed Mean

- For larger dataset.
- Excluding mean calculation by using assumption of mean like by median or mode.
- If value in decimal and you want whole number.

$A_x = 30$

$A_y = 33$

$d_x = X - 30$	$d_y = Y - 33$	$d_x \cdot d_y$	d_x^2	d_y^2
20	~~-21~~	420	400	441
-10	~~-8~~	80	100	64
0	0	0	0	0
10	12	120	100	144
20	22	440	400	484

$$r = \frac{1060}{\sqrt{1000 * 1133}} = 0.996$$

$dx = X - 30 \Rightarrow -20, -10, 0, 10, 20$

$dy = Y - 34 \Rightarrow -22, -9, -1, 11, 21$

$dx.dy \Rightarrow +440, 90, 0, 110, +420$

$dx^2 \Rightarrow 400, 100, 0, 100, 400$

$dy^2 \Rightarrow 484, 81, 1, 121, 441$

$$r = \frac{\Sigma dx.dy}{\sqrt{\Sigma dx^2 . \Sigma dy}}$$

$$= \frac{1060}{\sqrt{1000 \times 1128}} \Rightarrow \frac{1060}{1062.07}$$

$$= 0.998$$

Correlation.

Actual mean method.

For small datasets due to complex calculations.

x	10	20	30	40	50
~~(y)~~	12	25	33	45	55

1- $\bar{x} = 30$ It can be decimal.

$\bar{y} = 34$ uniform

2- Compute deviations.

$dx = x - \bar{x}$