IC 102-Data Analysis and Interpretation

Milind Sohoni

Lecture 1: Basic Statistics

Course Conduct

- Timetable: Wednesday, Friday 11:00-12:30 and Thursday, 9:30-10:30
- Textbook: Sheldon Ross 3rd (Indian Edition)
- lots of case studies
- Use of **Scilab**, in class and away.
- Assessment: 2 quizes, midsem, endsem, 2 programming assignments.
  - Open notes, no photocopies and no texts. Only hand-written notes.
  - **honesty and hardwork**
- webpage : www.cse.iitb.ac.in/∼sohoni/IC102

What are the objectives

- The importance of data-how to use it, infer from it
- Developing models-probability theory
- Parameter estimation and Hypothesis testing-key attributes of real life systems
- Regression-fitting models to data
- Case Studies-from various real life situations

Philosophy and History

- Empiricism-that observations determine everything
  - others : intuition, aesthetics, holism, spiritualism
  - examples- music, arts, natural history and evolution, the brain, justice and so on
  - early users-taxation, agriculture, *astronomy*
- The loop-Observe, analyse, model, predict, use, and maybe intervene

$$\text{Statistics} \Rightarrow \text{Probability}$$

- Basic laws of physics-mechanics, electrostatistics, fluid mechanics, civil engineering and so on
- Social Sciences-Economics-The supply and demand curves

Name	A	B	C	D	E	F
Ability to Pay	4	4	5	6	8	8

The basic object

- Population: The table $\mathcal{X}$- of N rows and m columns.
- Sample: The table X- of n rows and m columns.
  - How is the sample to be selected?
- Each row is a *item*, and thus there are n items.
- Each column is an *attribute*-thus m attributes.
- Example-Groundwater data

Lat.	Long.	Date	Depth
18.67453	79.1411	12th January 2001	3.56

- The attribute values could be numbers of texts

Village	Gram Panch.	Tal.	Population	ST	Tanker
Golbhan	Dhamni	Shahpur	566	24	N
Dhamni	Dhamni	Shahpur	376	376	Y

- Textual attributes are important, for they reveal the structure for analysis.

The simplest statistics

- The simplest dataset $X = \{x_1, \ldots, x_n\}$

Student	Marks
i	x_i

- *mean, mode median*
  - mean=$\overline{x} = (\sum_{i=1}^{n} x_i)/n$-average behaviour
  - median is x_M such that as many items below x_M asabove.
- Sample Variance $s^2 = (\sum_{i=1}^{n}(x_i - \overline{x})^2)/(n-1)$
- Sample *Standard Deviation* s.

Variance : the first measure of uncertainty.

- Suppose that the inter-arrival times are 10 minutes on the average but with a standard deviation of 3 minutes.
- Indian life expectancy is 64 years with $s = 15$.
- Rainfall is unchanged but variance has increased.

Chebyshev

- Two-sided: $N(S_k) =$ number of items such that $|x_i - \overline{x}| < ks$

$$\frac{N(S_k)}{n} \geq 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k^2}$$

Proof:

$$\begin{aligned}(n-1)s^2 &= \sum_i (x_i - \overline{x})^2 \\ &\geq \sum_{i:|x_i - \overline{x}| > ks} (x_i - \overline{x})^2 \\ &\geq (n - N(S_k))k^2 s^2\end{aligned}$$

$$\Rightarrow \frac{n-1}{nk^2} \geq (1 - \frac{N(S_k)}{n})$$

- One-sided: $N(k) =$ number of items such that $x_i - \overline{x} \geq ks$

$$\frac{N(k)}{n} \leq \frac{1}{1+k^2}$$

Limits on how 'far' data points can be from mean. Usually data sets are more bunched than Chebyshev.

Paired data-sets and Correlation

Student	Sem 1 SPI	Sem 2 SPI
i	x_i	y_i

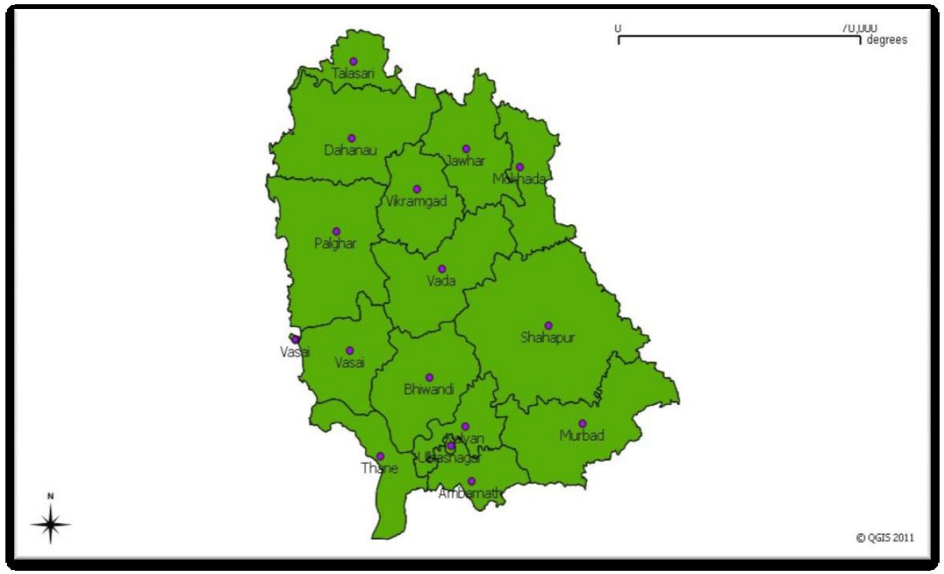
Correlation

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \overline{x})^2)(\sum_{i=1}^{n}(y_i - \overline{y})^2)}}$$

- $-1 \leq r \leq 1$
- if $y_i = ax_i + b$ then $r = sign(b)$
- $r(x, y) = r(ax + b, cy + d)$

Other analyses-The Thane DW Case-Study-I

- Grouping by textual attributes can be very important.



Area	9000 sq km.
Pop. (Rural) la.	81 (23)
Taluka (Tribal)	15 (5)
Habitats (GPs)	8000 (900)
Cities (Munci.)	37 (12)

And a severe drinking water probelm in 180 habitations.

How is the political structure?

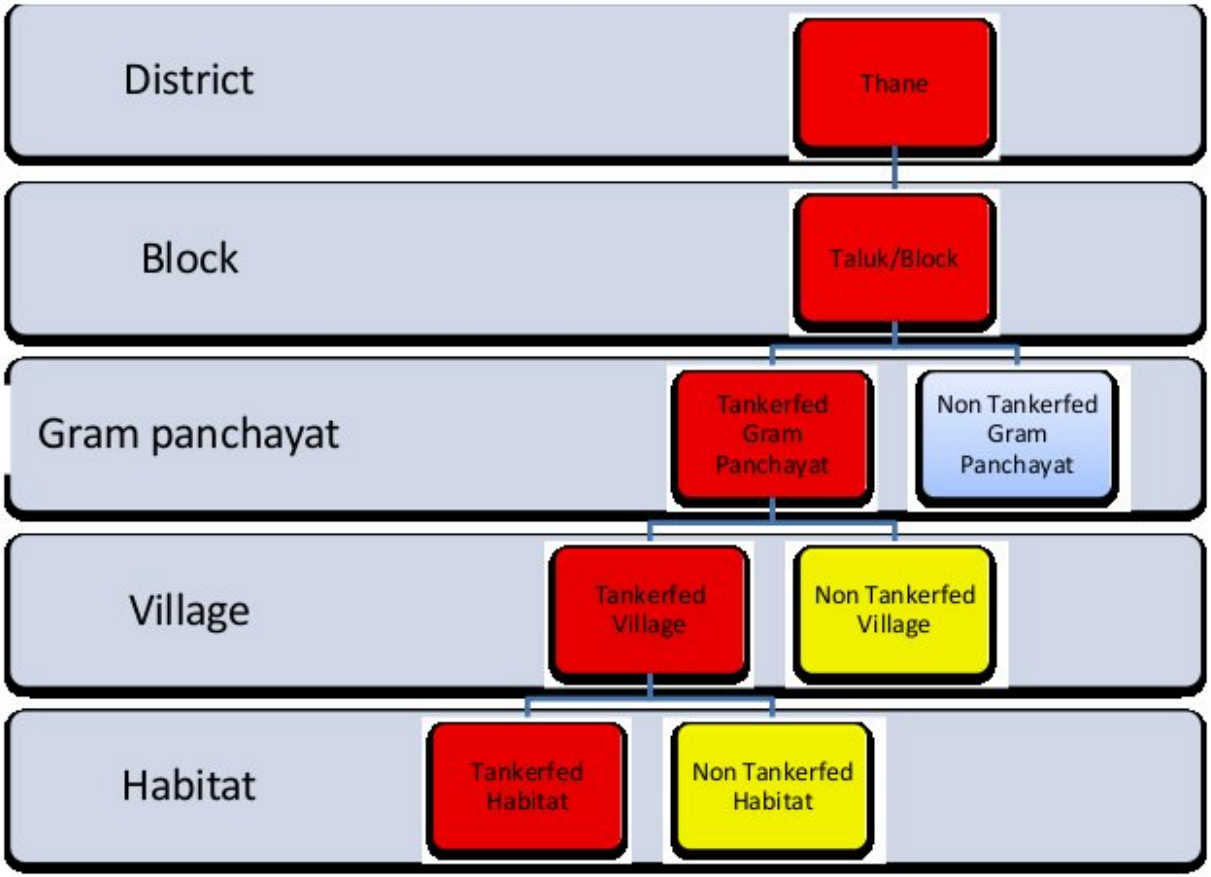
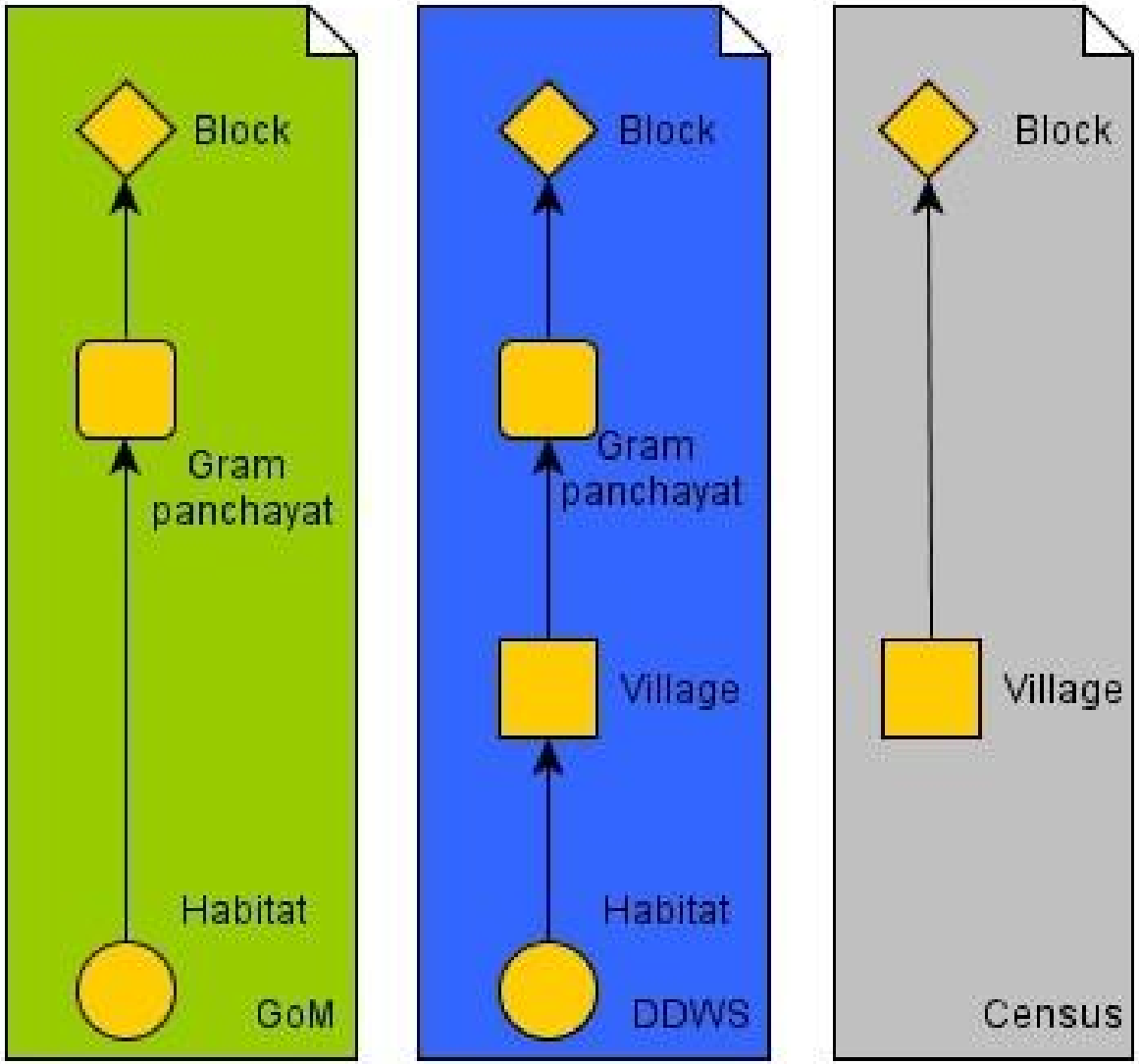


Figure 1:3: Data analysis levels

How is the data

61	AWALE	AWALE	AMBEPADA	2	0	0	0	0	1	0
62	AWALE	AWALE	AWALE	2	1	0	1	0	1	0
63	AWALE	AWALE	BORICHAPADA	1	0	0	0	0	1	1
64	AWALE	AWALE	JAMBHALIPADA	2	0	0	0	0	1	0
65	AWALE	AWALE	PATILPADA	2	0	0	0	0	1	0
66	AWALE	AWALE	PATRAYACHAPADA	1	0	0	0	0	1	1
67	AWALE	AWALE	WEDUCHAPADA	2	0	0	0	0	1	0
68	AWALE	CHANDROTI	CHANDROTI	2	1	0	0	0	0	0
69	AWALE	CHANDROTI	KATEKUIPADA	2	0	0	0	0	0	0
70	AWALE	CHANDROTI	KATKARIPADA (THA	2	0	0	0	0	0	0
71	AWALE	CHANDROTI	SHEKTYACHAPADA	2	0	0	0	0	0	0
72	AWALE	KARADE	KARADE	2	1	0	0	0	0	0
73	AWALE	MAHULI	KHARMEPADA	2	0	0	0	0	0	0
74	AWALE	MAHULI	MAHULI	1	1	1	0	0	0	0
75	AWALE	MAHULI	SUTARPADA	2	0	0	0	0	0	0
76	AWALE	MAMNOLI	DHAKANEPADA	2	0	0	0	0	0	0
77	AWALE	MAMNOLI	MAMNOLI	2	1	0	0	0	0	0
78	AWALE	MAMNOLI	PACHOLKARPADA	2	0	0	0	0	0	0
79	TALWADE	CHONDE KH)	CHONDE KH)	2	1	0	0	0	0	0
80	TALWADE	GANDHULWA	GANDHULWAD	1	1	1	0	0	0	0
81	TALWADE	HINGLUD	HINGLUD	1	1	1	0	0	0	0
82	TALWADE	ROADVAHAL	ROADVAHAL	1	1	1	0	0	0	0
83	TALWADE	TALWADE	TALWADE	1	1	1	1	1	1	1
84	DHASAI	DHASAI	DHASAI	2	1	0	1	0	1	0
85	DHASAI	DHASAI	KATKARIWADI	2	0	0	0	0	1	0
86	DHASAI	DHASAI	SAKHARWADI	2	0	0	0	0	1	0
87	DHASAI	MANGAON (N	MANGAON (THALK'	2	0	0	0	0	0	0
88	DHASAI	SHIVANERI	FARDEPADA	2	0	0	0	0	0	0
89	DHASAI	SHIVANERI	SHIVANERI	1	1	1	0	0	0	0
90	MANEKHIND	ADIVALI	ADIVALI	2	1	0	0	0	0	0
91	MANEKHIND	ADIVALI	ADIWALIPADA	2	0	0	0	0	0	0
92	MANEKHIND	ADIVALI	PAYERWADI	2	0	0	0	0	0	0
93	MANEKHIND	AMBEKHOR	AMBEKHOR	1	1	1	0	0	0	0
94	MANEKHIND	AMBEKHOR	BHALYACHIWADI	2	0	0	0	0	0	0
95	MANEKHIND	AMBEKHOR	KAVTEWADI	2	0	0	0	0	0	0
96	MANEKHIND	AMBEKHOR	KUBHAICHIWADI	2	0	0	0	0	0	0
97	MANEKHIND	AMBEKHOR	SAKHARWADI	2	0	0	0	0	0	0
98	MANEKHIND	ASHTE	ASHTE	2	1	0	0	0	0	0
99	MANEKHIND	MANEKHIND	MANEKHIND	2	1	0	1	0	1	0
100	MANJARE	BHINAR	BHINAR	2	1	0	0	0	0	0
101	MANJARE	MALAD	MALAD	2	1	0	0	0	0	0
102	MANJARE	MANGAON (N	MANGAON (N.V.)	2	1	0	0	0	0	0
103	MANJARE	MANJARE	MANJARE	1	1	1	1	1	1	1
104	MANJARE	MANJARE	NIBHALPADA	1	0	0	0	0	1	1
105	MANJARE	MANJARE	PANDHARICHAPAD	1	0	0	0	0	1	1
106	MANJARE	TEMBHURLI	CHANDRICHAPADA	2	0	0	0	0	0	0
107	MANJARE	TEMBHURLI	KATKARIWADI	2	0	0	0	0	0	0
108	MANJARE	TEMBHURLI	TEMBHURLI	1	1	1	0	0	0	0
109	MANJARE	TEMBHURLI	TORANPADA	2	0	0	0	0	0	0
110	VASHALA (BK	VASHALA BK	CHARANWADI	1	0	0	0	0	0	0
111	VASHALA (BK	VASHALA BK	KOLIPADA	2	0	0	0	0	0	0
112	VASHALA (BK	VASHALA BK	RAICHEWADI	1	0	0	0	0	0	0

How are the data-bases



Some Raw Statistics

Taluka	Villages	Tank.	Frac.	Area	T. Area	Frac.
Jawhar	61	14	0.23	609	96	0.16
Mokhada	79	25	0.32	494	277	0.56
Murbad	199	10	0.05	913	688	0.75
Shahpur	224	34	0.16	1604	463	0.29

Social Indicators

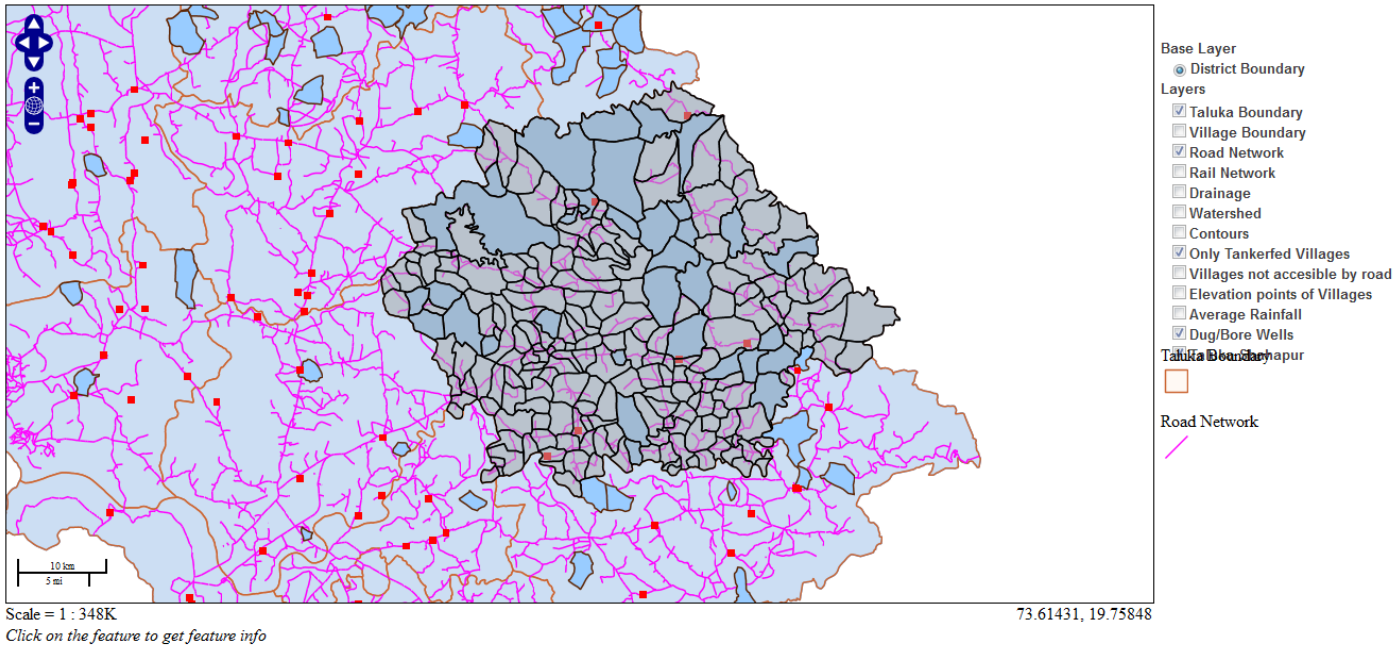
Fraction of Female Illiteracy

	Jawhar	Mokhada	Murbad	Shahpur
Tankerfed	0.78	0.69	0.64	0.67
Neighbors	0.76	0.70	0.61	0.56
Taluka	0.76	0.68	0.55	0.55

Fraction of ST population.

	Jawhar	Mokhada	Murbad	Shahpur
Tankerfed	0.97	0.93	0.74	0.62
Neighbors	0.99	0.97	0.32	0.42
Taluka	0.97	0.91	0.24	0.35

GIS Shahpur closeup



Elevations

	Jawhar	Mokhada	Murbad	Shahpur
Tankerfed	344	361	123	197
Taluka	320	350	126	132

Thanks