Descriptive Statistics

Fall 2024

Instructor:

Sunita Sarawagi

Thanks to Prof Ajit Rajawade for the initial version of the slides

Terminology

- **Population**: The collection of all elements which we wish to study, example: data about occurrence of tuberculosis all over the world
- In this case, “population” refers to the set of people in the entire world.
- The population is often too large to examine/study.
- So we study a subset of the population – called as a **sample**.
- In an experiment, we basically collect **values** for one or more **attributes or variables** of each member of the sample.

Examples of samples

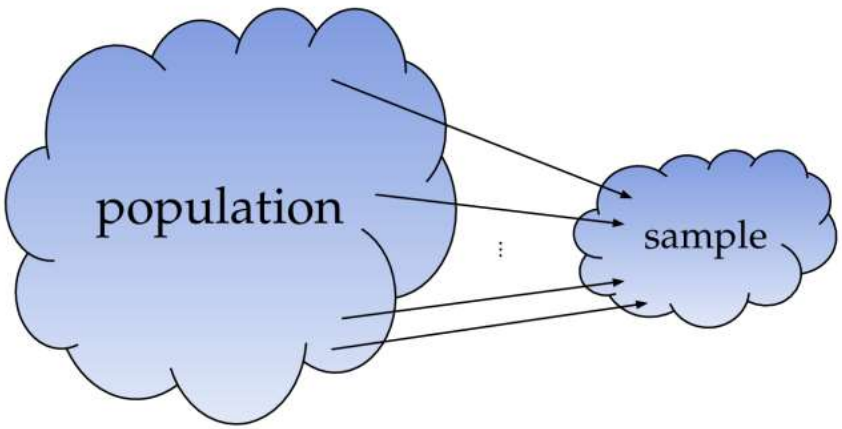
variable

index	username	country	age	ezlvl	time	points	finished
0	mary	us	38	0	124.94	418	0
1	jane	ca	21	0	331.64	1149	1
2	emil	fr	52	1	324.61	1321	1
3	ivan	ca	50	1	39.51	226	0
4	hasan	tr	26	1	253.19	815	0
5	jordan	us	45	0	28.49	206	0
6	sanjay	ca	27	1	585.88	2344	1
7	lena	uk	23	0	408.76	1745	1
8	shuo	cn	24	1	194.77	1043	0
9	r0byn	us	59	0	255.55	1102	0
10	anna	pl	18	0	303.66	1209	1
11	joro	bg	22	1	381.97	1491	1

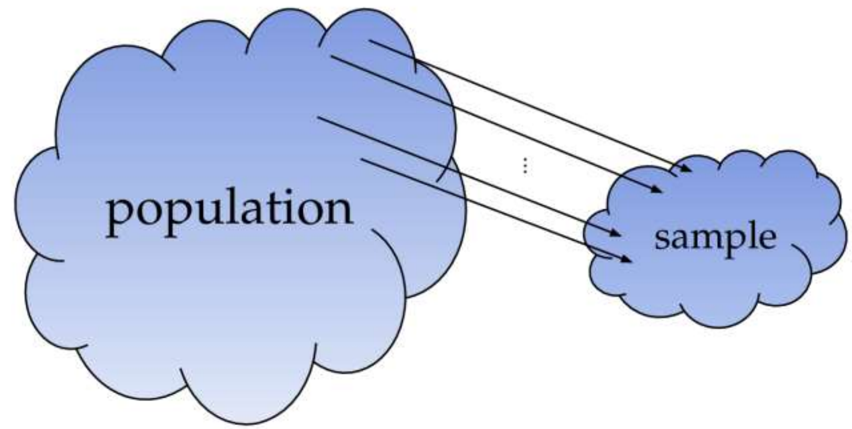
observation

Table 1.1: A data table that contains observations of seven variables for 12 players of a computer game. Each row in this table corresponds to one player. Each column corresponds to one characteristic that was measured for all the players.

Population and Samples



(a) Representative sample selection



(b) Biased sample selection

Data Representation and Visualization

Need for data visualization

- The raw dataset or tables may be too large. Cannot make sense of the data just by inspecting raw table of numbers.

- Even if data is not too large, patterns emerge sometimes only under right type of visualization.

Outline

- Visualizing values of each variable separately
- Visualizing pairs of variables.
- Multi-dimensional data

Terminology

- **Discrete data:** Data whose values are restricted to a finite or countably infinite set. Eg: letter grades at IITB, genders, marital status (single, married, divorced), income brackets in India for tax purposes

- **Continuous data:** Data whose values belong to an uncountably infinite set (Eg: a person's height, temperature of a place, speed of a car at a time instant).

Raw data

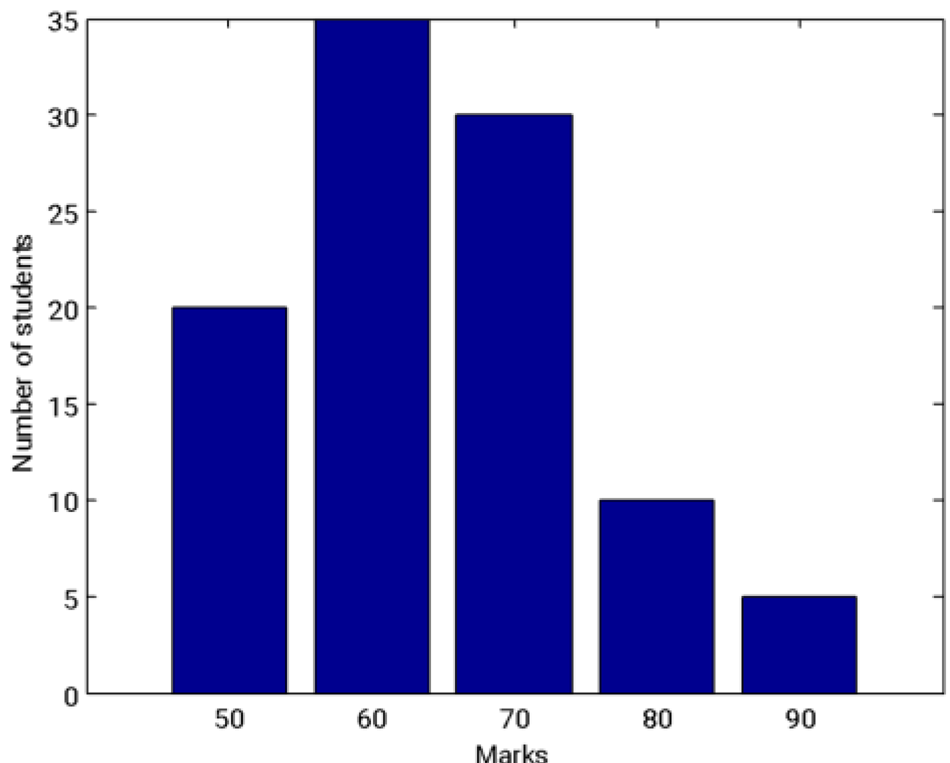
- Example: Country of winners of any competition
- Example: Grades of students in CS 215

Frequency Tables

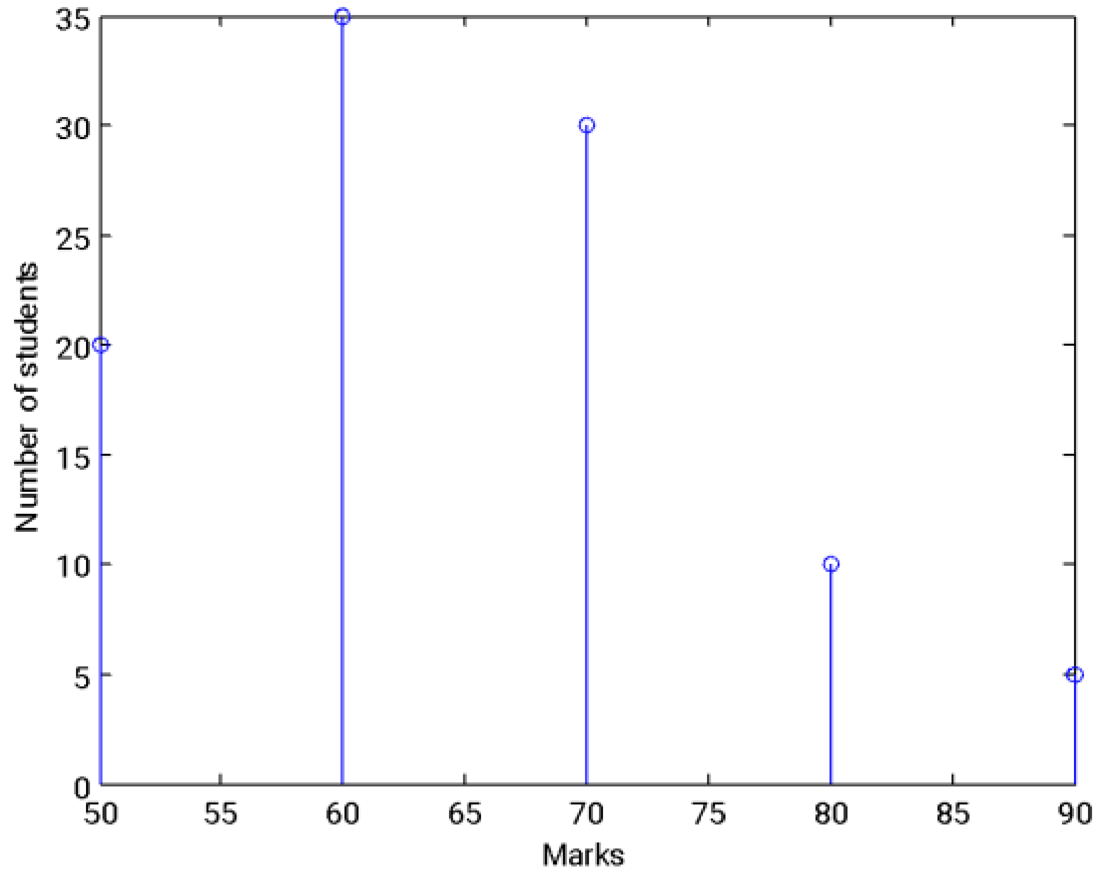
- The frequency table can be visualized using a **line graph** or a **bar graph** or a **frequency polygon**.

Grade	Number of students
AA	5
AB	10
BB	30
BC	35
CC	20

A **bar graph** plots the distinct data values on the X axis and their frequency on the Y axis by means of the height of a thick vertical bar!

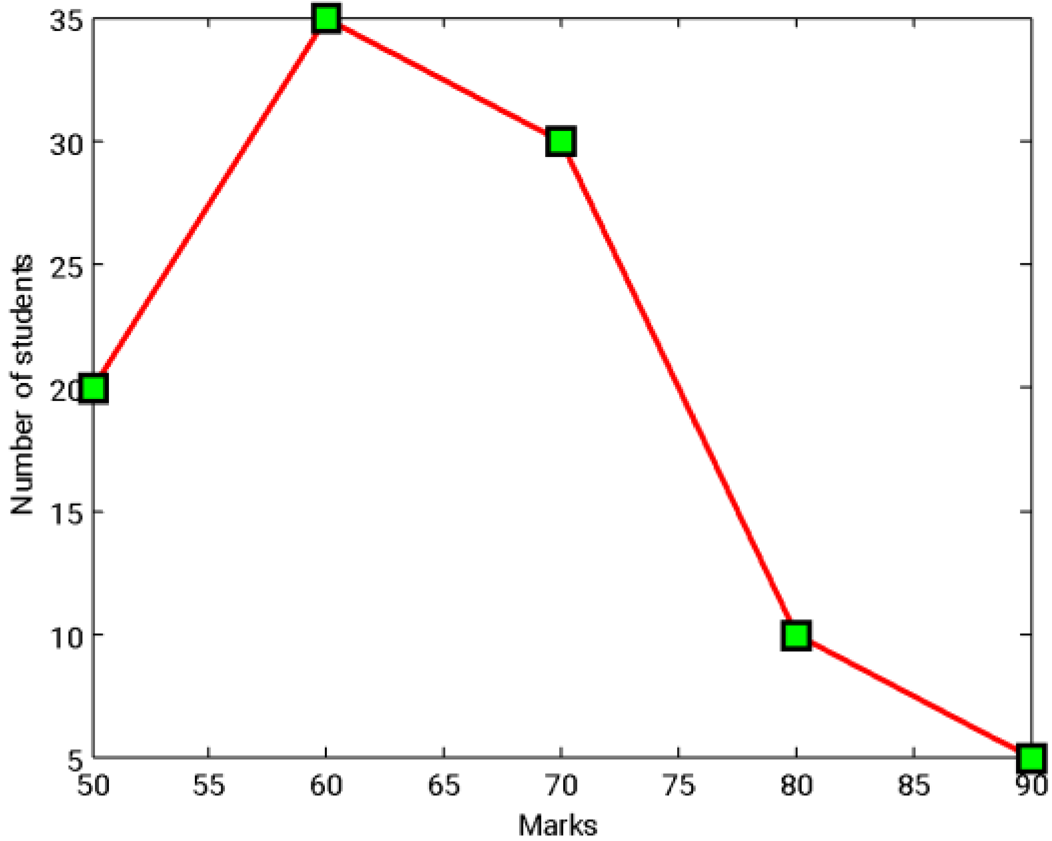


Grade	Number of students
AA	5
AB	10
BB	30
BC	35
CC	20



A **line diagram** plots the distinct data values on the X axis and their frequency on the Y axis by means of the height of a vertical line!

Grade	Number of students
AA	5
AB	10
BB	30
BC	35
CC	20



A **frequency polygon** plots the frequency of each data value on the Y axis, and connects consecutive plotted points by means of a line.

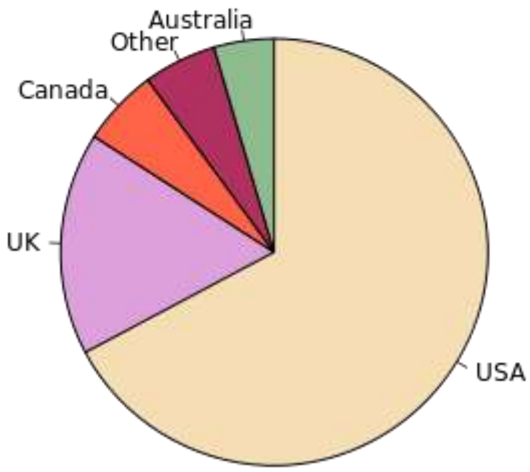
Relative frequency tables

- Sometimes the actual frequencies are not important.
- We may be interested only in the *percentage* or *fraction* of those frequencies for each data value – i.e. *relative frequencies*.

Grade	Fraction of number of students
AA	0.05
AB	0.10
BB	0.30
BC	0.35
CC	0.20

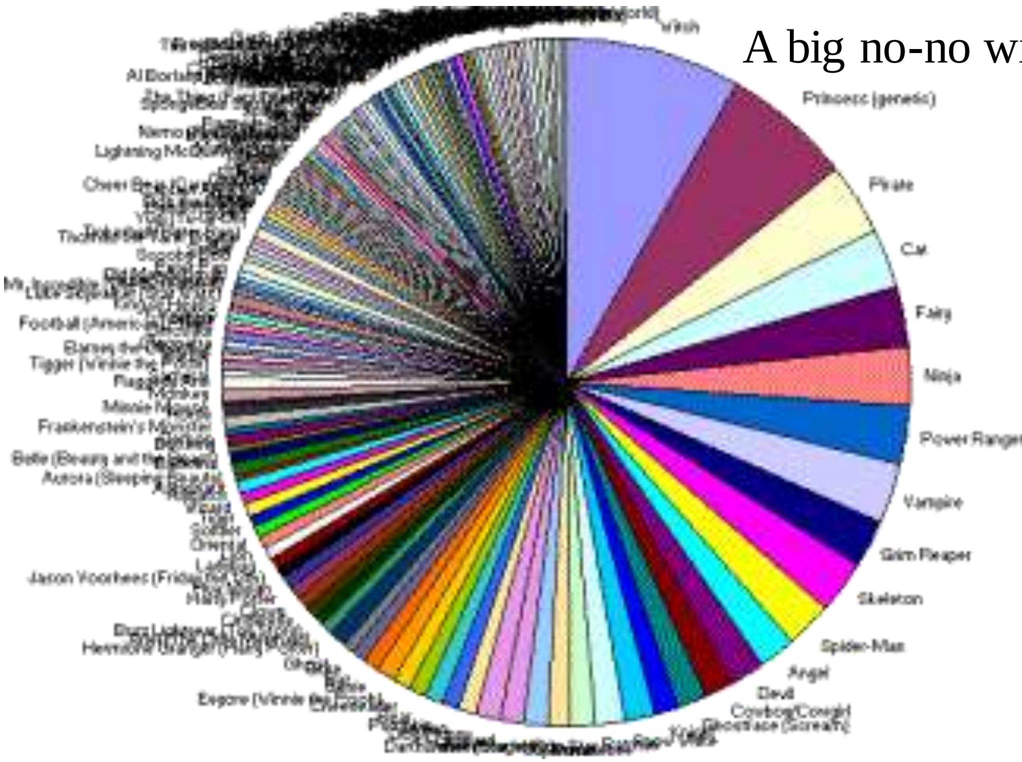
Pie charts

- For a small number of distinct data values which are non-numerical, one can use a **pie-chart** (it can also be used for numerical values).
- It consists of a circle divided into sectors corresponding to each data value.
- The area of each sector = relative frequency for that data value.



Population of native English speakers: https://en.wikipedia.org/wiki/Pie_chart

Pie charts can be confusing



A big no-no with too many categories.

http://stephenturbek.com/articles/2009/06/better-charts-from-simple-questions.html

Dealing with continuous data

- Example: temperature of a place at a time instant, speed of a car at a given time instant, weight or height of an animal, etc.

- The raw data: marks in final exams.

Table 2.3 Life in Hours of 200 Incandescent Lamps.

Item Lifetimes									
1067	919	1196	785	1126	936	918	1156	920	948
855	1092	1162	1170	929	950	905	972	1035	1045
1157	1195	1195	1340	1122	938	970	1237	956	1102
1022	978	832	1009	1157	1151	1009	765	958	902
923	1333	811	1217	1085	896	958	1311	1037	702
521	933	928	1153	946	858	1071	1069	830	1063
930	807	954	1063	1002	909	1077	1021	1062	1157
999	932	1035	944	1049	940	1122	1115	833	1320
901	1324	818	1250	1203	1078	890	1303	1011	1102
996	780	900	1106	704	621	854	1178	1138	951
1187	1067	1118	1037	958	760	1101	949	992	966
824	653	980	935	878	934	910	1058	730	980
844	814	1103	1000	788	1143	935	1069	1170	1067
1037	1151	863	990	1035	1112	931	970	932	904
1026	1147	883	867	990	1258	1192	922	1150	1091
1039	1083	1040	1289	699	1083	880	1029	658	912
1023	984	856	924	801	1122	1292	1116	880	1173
1134	932	938	1078	1180	1106	1184	954	824	529
998	996	1133	765	775	1105	1081	1171	705	1425
610	916	1001	895	709	860	1110	1149	972	1002

Visualizing numerical data

- Reduce to a known problem
  - Group into bins/intervals
  - Count number in each bin.
  - Draw histogram

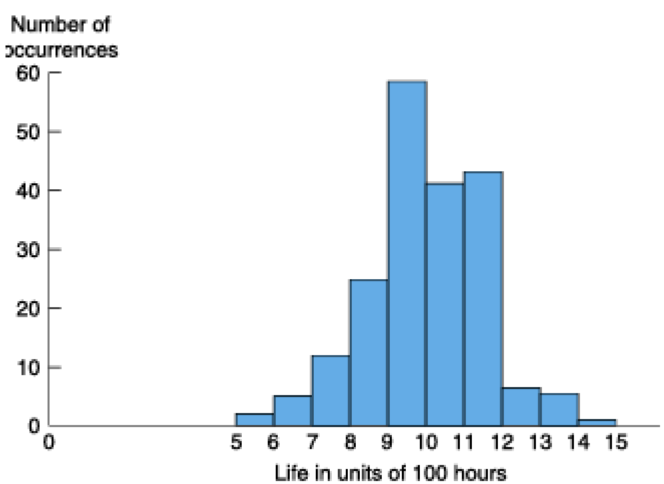


Table 2.4 A Class Frequency Table.

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1

Dealing with continuous data

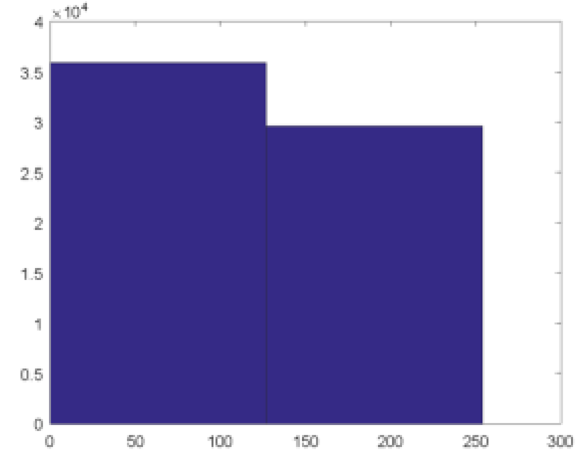
- Let the sample points be $\{x_i\}$, 1 <= i <= N.
- Let there be some K (K << N) bins, where the j^{th} bin has interval $[a_j, b_j)$.
- Thus frequency f_j for the j^{th} bin is defined as follows:

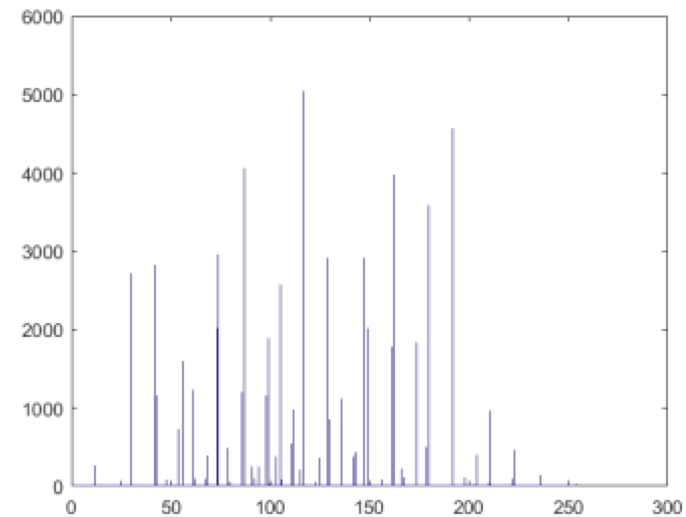
$$f_j = |\{x_i : a_j \leq x_i < b_j, 1 \leq i \leq N\}|$$

- Such frequency tables are also called **histograms** and they can also be used to store relative frequency instead of frequency.

The histogram binning problem

- If you have too few bins (each bin is very wide), there is very little idea you get about the data distribution from the histogram.
- Extreme: only one bin to represent whole data
- If you have many bins (all will be narrow), then there are very points falling into each bin. Again there is very little idea you get about the data distribution from the histogram.
- Extreme: One bin for each distinct value



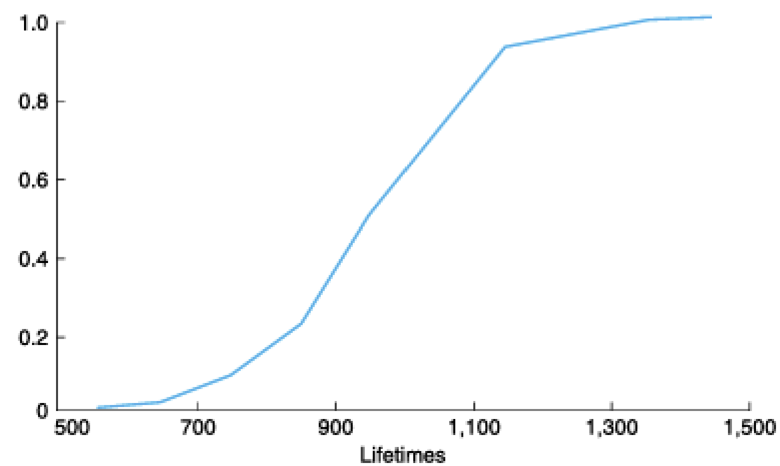


Cumulative frequency plot

The **cumulative** (relative) **frequency plot** tells you the (proportion) number of sample points whose value is *less than or equal to* a given data value.

Table 2.4 A Class Frequency Table.

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1



Summarizing Data

08	02	22	97	38	15	00	40	00	75	04	05	07	78	52	12	50	77	91	08
49	49	99	40	17	81	18	57	60	87	17	40	98	43	69	48	04	56	62	00
81	49	31	73	55	79	14	29	93	71	40	67	53	88	30	03	49	13	36	65
52	70	95	23	04	60	11	42	69	24	68	56	01	32	56	71	37	02	36	91
22	31	16	71	51	67	63	89	41	92	36	54	22	40	40	28	66	33	13	80
24	47	32	60	99	03	45	02	44	75	33	53	78	36	84	20	35	17	12	50
32	98	81	28	64	23	67	10	26	38	40	67	59	54	70	66	18	38	64	70
67	26	20	68	02	62	12	20	95	63	94	39	63	08	40	91	66	49	94	21
24	55	58	05	66	73	99	26	97	17	78	78	96	83	14	88	34	89	63	72
21	36	23	09	75	00	76	44	20	45	35	14	00	61	33	97	34	31	33	95
78	17	53	28	22	75	31	67	15	94	03	80	04	62	16	14	09	53	56	92
16	39	05	42	96	35	31	47	55	58	88	24	00	17	54	24	36	29	85	57
86	56	00	48	35	71	89	07	05	44	44	37	44	60	21	58	51	54	17	58
19	80	81	68	05	94	47	69	28	73	92	13	86	52	17	77	04	89	55	40
04	52	08	83	97	35	99	16	07	97	57	32	16	26	26	79	33	27	98	66
88	36	68	87	57	62	20	72	03	46	33	67	46	55	12	32	63	93	53	69
04	42	16	73	38	25	39	11	24	94	72	18	08	46	29	32	40	62	76	36
20	69	36	41	72	30	23	88	34	62	99	69	82	67	59	85	74	04	36	16
20	73	35	29	78	31	90	01	74	31	49	71	48	86	81	16	23	57	05	54
01	70	54	71	83	51	54	69	16	92	33	48	61	43	52	01	89	19	67	48

Summarizing a sample-set

- There are some values that can be considered "representative" of the entire sample-set. Such values are called as a "statistic".
- The most common statistic is the sample (arithmetic) **mean**:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- It is basically what is commonly regarded as "average value".

Summarizing a sample-set

- Another common statistic is the sample **median**, which is the “middle value”.

- We sort the data array **A** from smallest to largest. If N is odd, then the median is the value at the $(N+1)/2$ position in the sorted array.

- If N is even, the median can take any value in the interval $(A[N/2],A[N/2+1])$ – why?

Properties of the mean and median

- Consider each sample point x_i were replaced by $ax_i + b$ for some constants a and b.

- What happens to the mean? What happens to the median?

- Consider each sample point x_i were replaced by its square.

- What happens to the mean? What happens to the median?

Properties of the mean and median

- ***Question:*** Consider a set of sample points x_1, x_2, …, x_N. For what value *y*, is the sum total of the **squared** difference with every sample point, the least? That is, what is:

$$\arg\min_y \sum_{i=1}^{N} (y - x_i)^2 \text{ ?}$$

Total squared deviation
(or total squared loss)

Answer: mean

- ***Question:*** For what value *y*, is the sum total of the **absolute** difference with every sample point, the least? That is, what is:

$$\arg\min_y \sum_{i=1}^{N} | y - x_i | \text{?}$$

Total absolute deviation
(or total absolute loss)

Answer: median

Proof that mean minimizes square deviation

Proof that median minimize absolute deviation

Properties of the mean and median

- The mean need not be a member of the original sample-set.
- The median is always a member of the original sample-set if N is odd.
- The median is not unique and will not be a member of the set if N is even.

Properties of the mean and median

- Consider a set of sample points $x_1, x_2, \ldots, x_N$. Let us say that some of these values get grossly corrupted.
- What happens to the mean?
- What happens to the median?

Example

- Let A ={1,2,3,4,6}
- Mean (A) = 3.2, median (A) = 3
- Now consider A = {1,2,3,4,20}
- Mean (A) = 6, median(A) = 3.

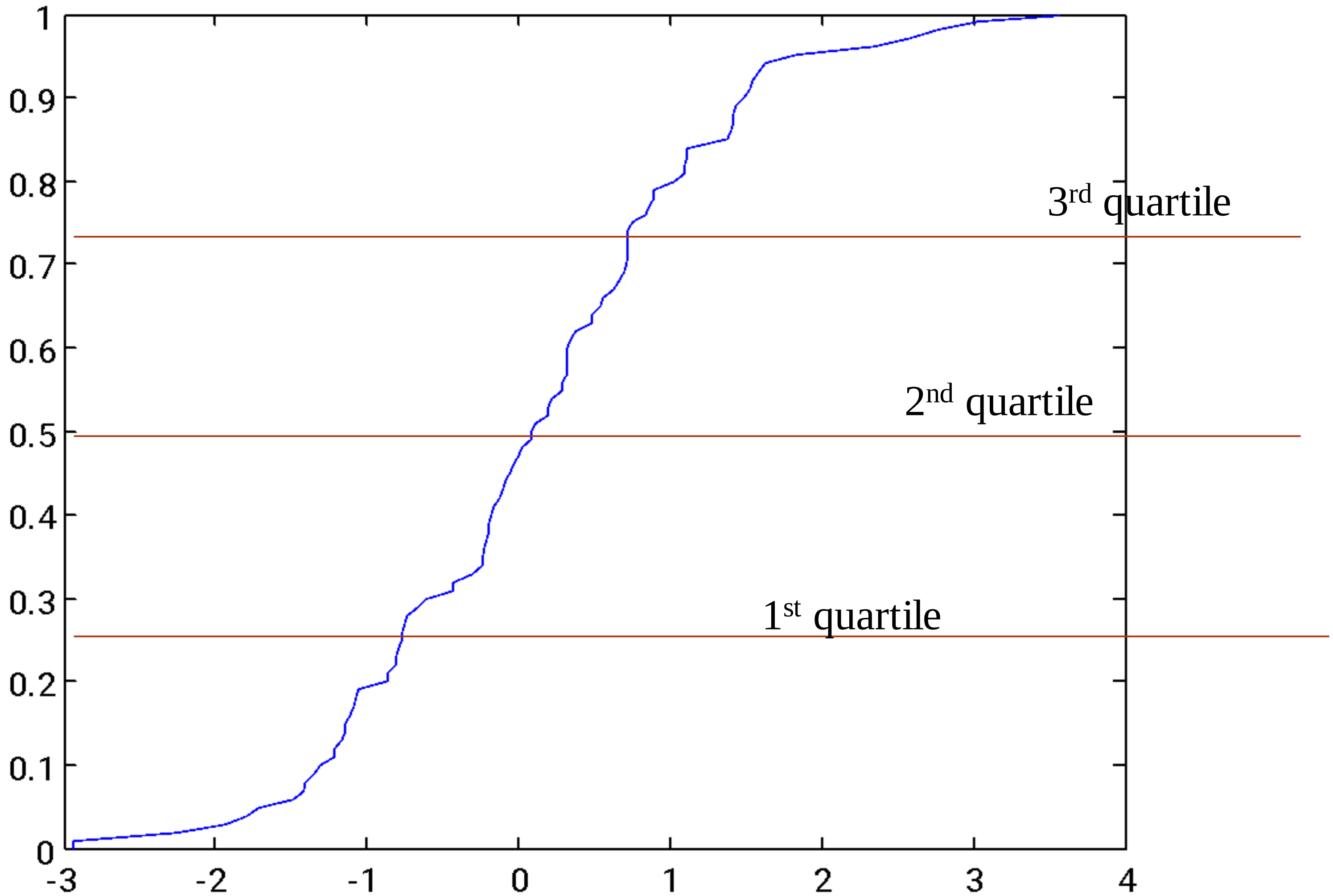
Percentiles

- The sample 100*p* percentile ($0 \leq p \leq 1$) is defined as the data value *y* such that 100*p*% of the data have a value less than or equal to *y*, and 100(1-*p*)% of the data have a larger value.

- For a data set with *n* sample points, the sample 100*p* percentile is that value such that at least *np* of the values are less than or equal to it. And at least *n*(1-*p*) of the values are greater than it.

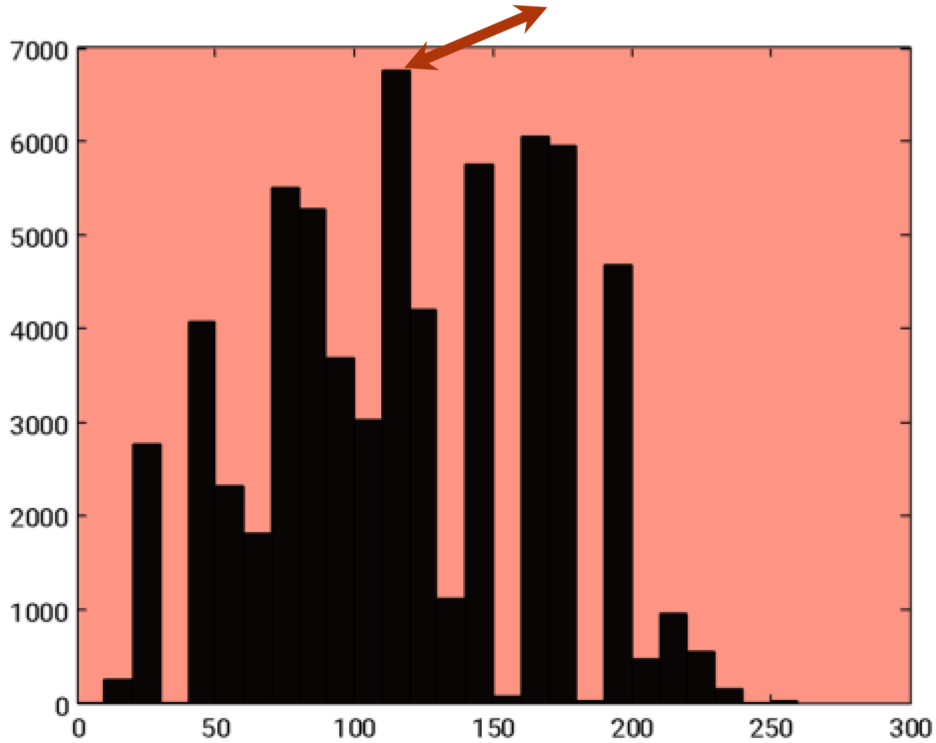
Quantiles

- The sample 25 percentile = first quartile.
- The sample 50 percentile = second quartile.
- The sample 75 percentile = third quartile.
- Quantiles can be inferred from the cumulative relative frequency plot (how?).
- Or by sorting the data values (how?).


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3
4
3rd quartile
2nd quartile
1st quartile

Mode

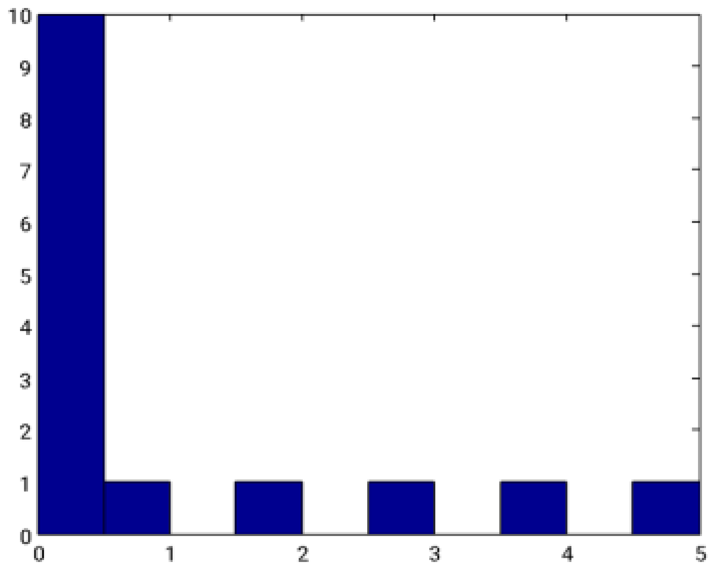
The value that occurs with the highest frequency is called the mode.

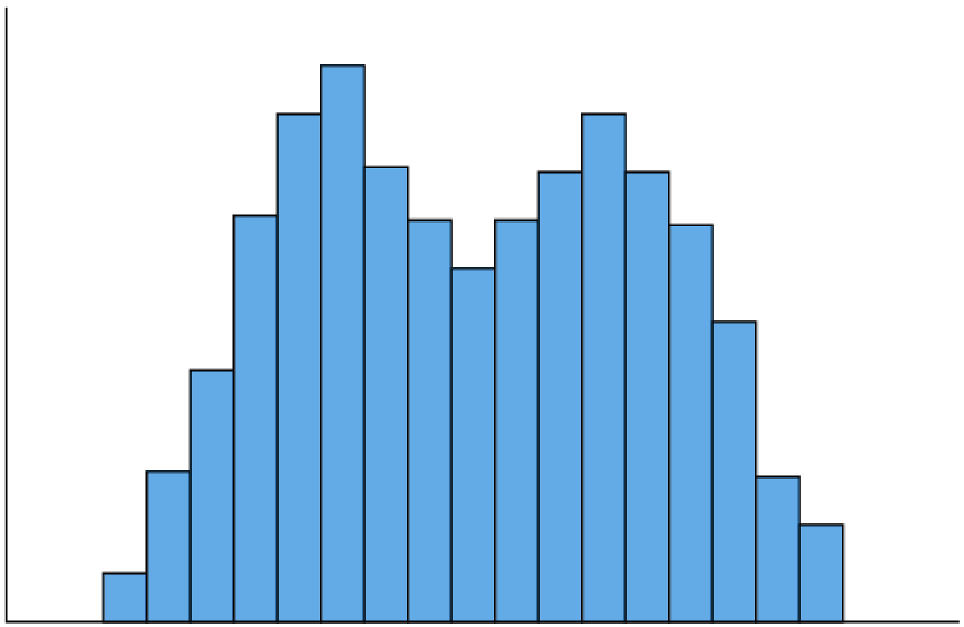


Mode

The mode may not be unique, in which case all the highest frequency values are called **modal values**.

Mode at 0





Variance and Standard deviation

- The **variance** is (approximately) the average value of the squared distance between the sample points and the sample mean. The formula is:

$$\text{variance} = s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{x} - x_i)^2$$

The division by N-1 instead of N is for a very technical reason which we will understand after many lectures. As such, the variance is computed usually when N is large so the numerical difference is not much.

- The variance measures the "spread of the data around the sample mean".
- Its positive square-root is called as the **standard deviation**.

Variance and Standard deviation: Properties

Consider each sample point x_i were replaced by $ax_i + b$ for some constants a and b. What happens to the standard deviation?

Chebyshev’s inequality

- Suppose you know the average marks for this course was 75 (out of 100). And that the variance of the marks was 25.
- Can you say something about how many students secured marks from 65 to 85?
- You obviously cannot predict the exact number – but you can say **something** about this number.
- That something is given by Chebyshev’s inequality.

Chebyshev's inequality: and Chebyshev

https://en.wikipedia.org/wiki/Pafnuty_Chebyshev

Russian mathematician:
Stellar contributions in probability and statistics, geometry, mechanics

Two-sided Chebyshev's inequality:

The proportion of sample points k or more than k $(k>0)$ standard deviations away from the sample mean is less than $1/k^2$.

Chebyshev's inequality: and Chebyshev

Two-sided Chebyshev's inequality:

The proportion of sample points k or more than k *($k>0$)* standard deviations away from the sample mean is less than or equal to $1/k^2$.

$$S_k = \{x_i : | x_i - \bar{x} | \geq k\sigma\}$$

$$\frac{|S_k|}{N} < \frac{1}{k^2}$$

Chebyshev's inequality

- Applying this inequality to the previous problem, we see that the fraction of students who got less than 65 or more than 85 marks is as follows:

$$S_k = \{x_i : |x_i - \bar{x}| \geq k\sigma\}$$

$$\frac{|S_k|}{N} \leq \frac{1}{k^2}$$

$$\bar{x} = 75$$

$$\sigma = 5$$

$$k = 2$$

$$\frac{|S_k|}{N} \leq \frac{1}{4}$$

- So the fraction of students who got from 65 to 85 is more than 1-0.25 = 0.75.

Chebyshev's inequality

1	Kerala	93.91
2	Lakshadweep	92.28
3	Mizoram	91.58
4	Tripura	87.75
5	Goa	87.40
6	Daman & Diu	87.07
7	Puducherry	86.55
8	Chandigarh	86.43
9	Delhi	86.34
10	Andaman & Nicobar Islands	86.27
11	Himachal Pradesh	83.78
12	Maharashtra	82.91

Mean = 87.69
Std. dev. = 3.306

Fraction of states with literacy rate in the range
$(\mu-1.5\sigma, \mu+1.5\sigma)$ is $11/12 \approx 91\%$

As predicted by Chebyshev's inequality, it is **at least**
$1-1/(1.5*1.5) \approx 0.55$

The bounds predicted by this inequality are loose – but they are correct!

https://en.wikipedia.org/wiki/Indian_states_ranking_by_literacy_rate

Proof of Chebyshev's inequality

One-sided Chebyshev's inequality

- Also called the Chebyshev-Cantelli inequality.

The proportion of sample points k or more than k $(k>0)$ standard deviations away from the sample mean **and greater than the sample mean** is less than or equal to $1/(1+k^2)$.

Notice: no absolute value!

$$S_k = \{x_i : x_i - \bar{x} \geq k\sigma\}$$

$$\frac{|S_k|}{N} \leq \frac{1}{1+k^2}$$

One-sided Chebyshev's inequality (Another form)

- Also called the Chebyshev-Cantelli inequality.

The proportion of sample points k or more than k *(k>0)* standard deviations away from the sample mean **and less than the sample mean** is less than or equal to $1/(1+k^2)$.

Notice: no absolute value!

$$S_k = \{x_i : x_i - \bar{x} \leq -k\sigma\}$$

$$\frac{|S_k|}{N} \leq \frac{1}{1+k^2}$$

Analyzing pairs of variables

Correlation between different data values

- Sometimes each sample-point can have a pair of attributes.

- And it may so happen that large values of the first attribute are accompanied with large (or small) values of the second attribute for a large number of sample-points.

Correlation between different data values

- Example 1: Populations with higher levels of fat intake show higher incidence of heart disease.
- Example 2: People with higher levels of education often have higher incomes.
- Example 3: Literacy Rate in India as a function of time?

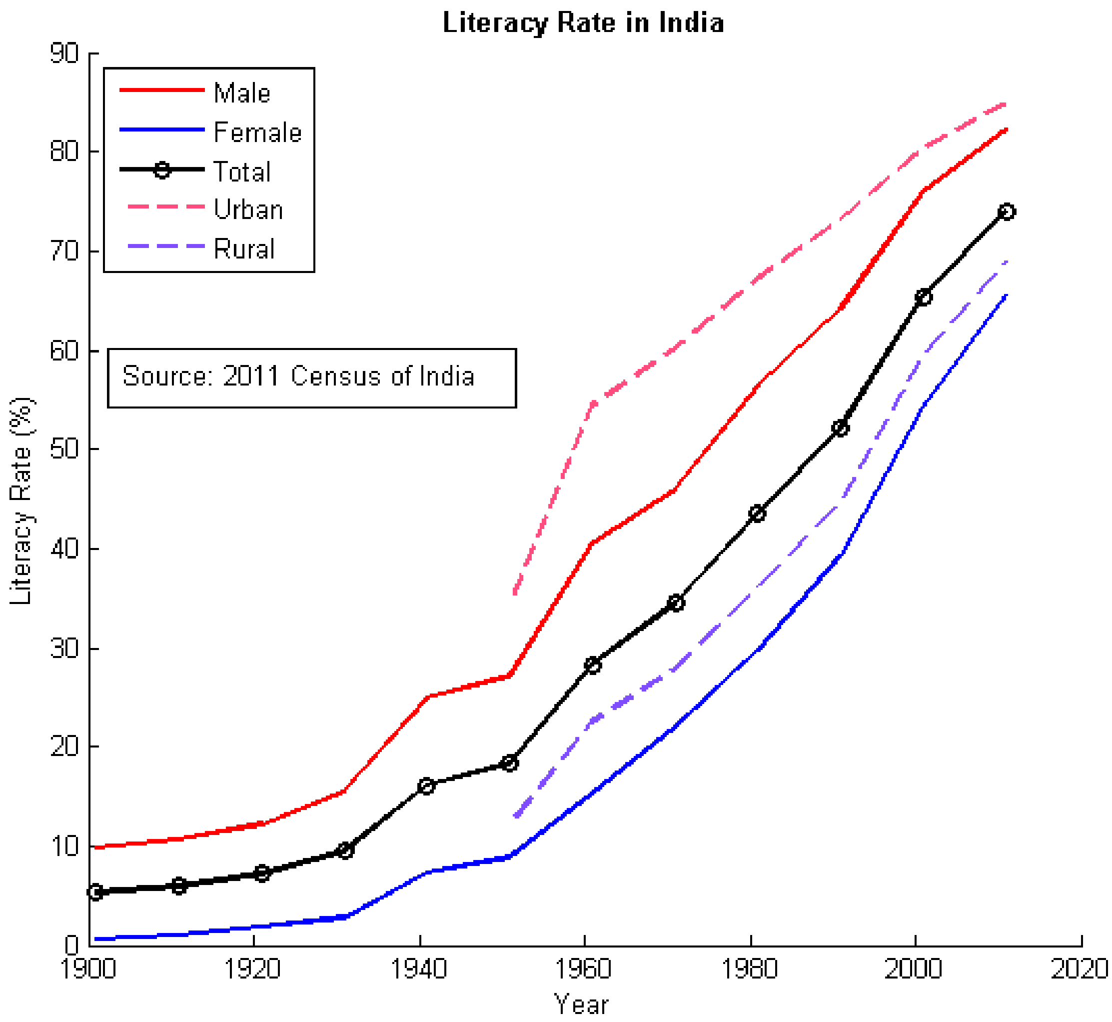


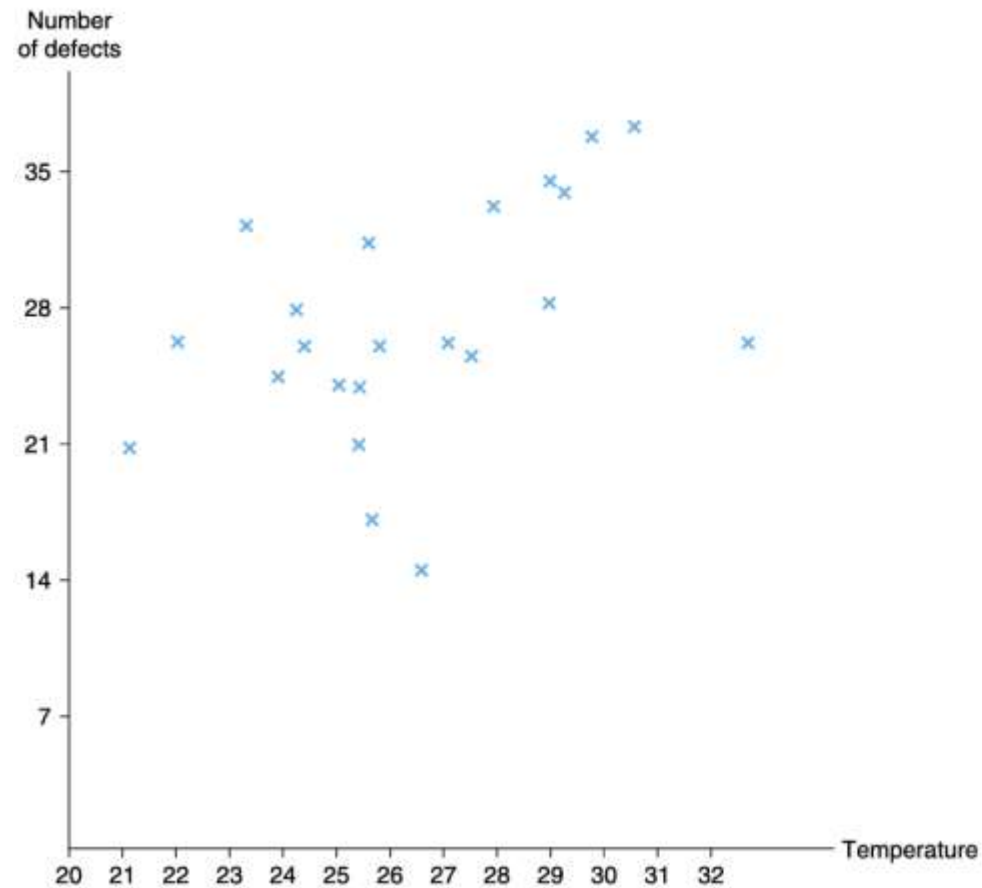
Image source

Visualizing such relationships?

- Can be done by means of a scatter plot
- X axis: values of attribute 1, Y axis: values of attribute 2
- Plot a marker at each such data point. The marker may be a small circle, a +, a *, and so on.

Table 2.8 Temperature and Defect Data.

Day	Temperature	Number of Defects
1	24.2	25
2	22.7	31
3	30.5	36
4	28.6	33
5	25.5	19
6	32.0	24
7	28.6	27
8	26.5	25
9	25.3	16
10	26.0	14
11	24.4	22
12	24.8	23
13	20.6	20
14	25.1	25
15	21.4	25
16	23.7	23
17	23.9	27
18	25.2	30
19	27.4	33
20	28.3	32
21	28.8	35
22	26.6	24



Correlation coefficient

- Let the sample-points be given as (x_i,y_i), 1 <= i <= N.
- Let the sample standard deviations be σ_x and σ_y, and the sample means be μ_x and μ_y.
- The **correlation-coefficient** is given as:

$$r(x,y) = \frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i-\mu_x)^2\sum_{i=1}^{N}(y_i-\mu_y)^2}} = \frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{(N-1)\sigma_x\sigma_y}$$

Correlation coefficient

- The correlation-coefficient is given as:

$$r(x,y) = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_x)^2 \sum_{i=1}^{N}(y_i - \mu_y)^2}} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{(N-1)\sigma_x \sigma_y}$$

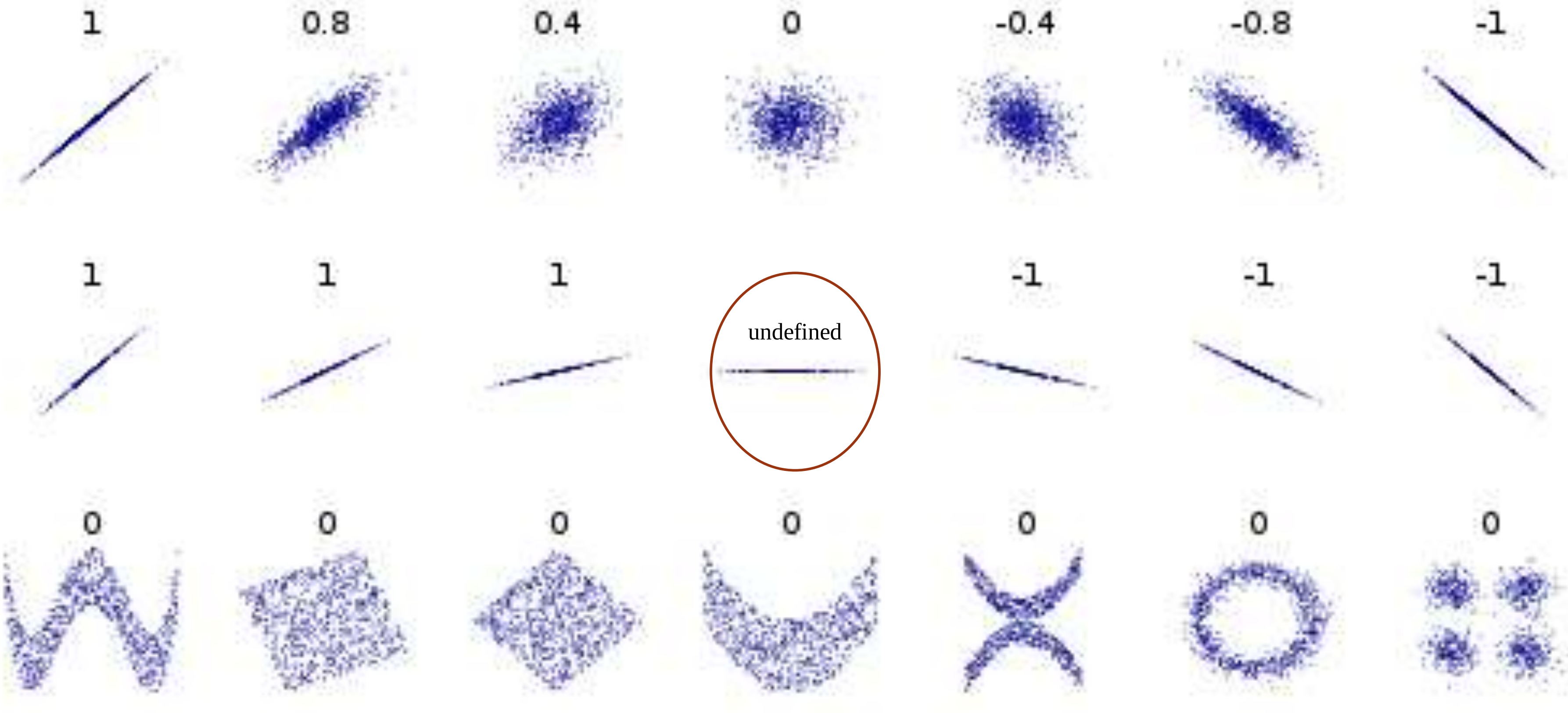
- $r > 0$ means the data are **positively correlated** (one attribute being higher implies the other is higher)
- $r < 0$ means the data are **negatively correlated** (one attribute being higher implies the other is lower)
- $r = 0$ means the data are **uncorrelated** (there is no such relationship!)
- r is **undefined** if the standard deviation of either ***x*** or ***y*** is 0.

Correlation coefficient: Properties

- The correlation-coefficient is given as:

$$r(x,y)=\frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i-\mu_x)^2\sum_{i=1}^{N}(y_i-\mu_y)^2}}=\frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{(N-1)\sigma_x\sigma_y}$$

- -1 <= *r* <= 1 always!



Correlation coefficient values for various toy datasets in 2D: for each dataset, a scatter plot is provided

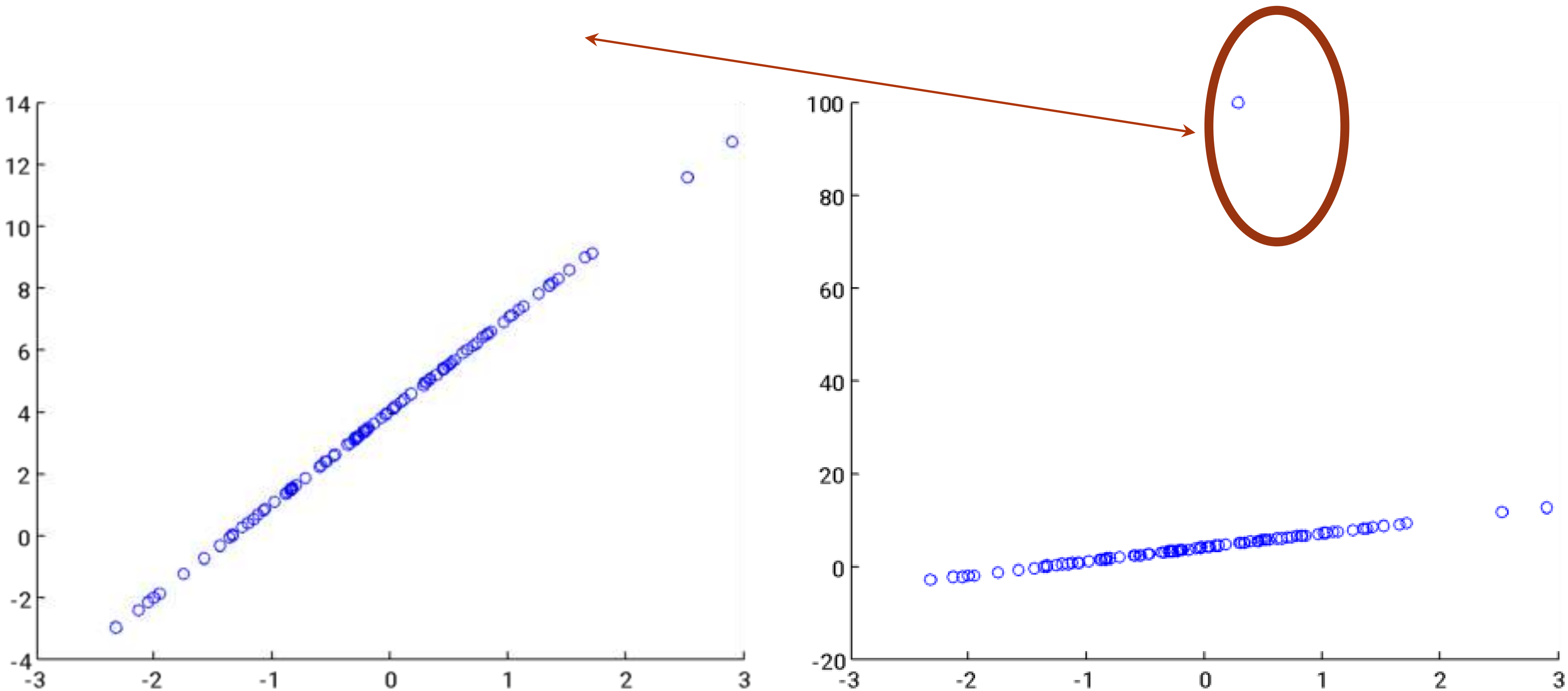
https://en.wikipedia.org/wiki/Correlation_and_dependence

Correlation coefficient: Properties

- In the following, we have a,b,c,d constant.
- If $y_i = a+bx_i$ where $b > 0$, then $r(x,y) = 1$.
- If $y_i = a+bx_i$ where $b < 0$, then $r(x,y) = -1$.
- If r is the correlation coefficient of data pairs as (x_i,y_i), $1 <= i <= N$, then it is also the correlation coefficient of data pairs $(b+ax_i,d+cy_i)$ when a and c have the same sign.

Correlation coefficient: a word of caution

- Sensitive to outliers!



r = 1

r = 0.33

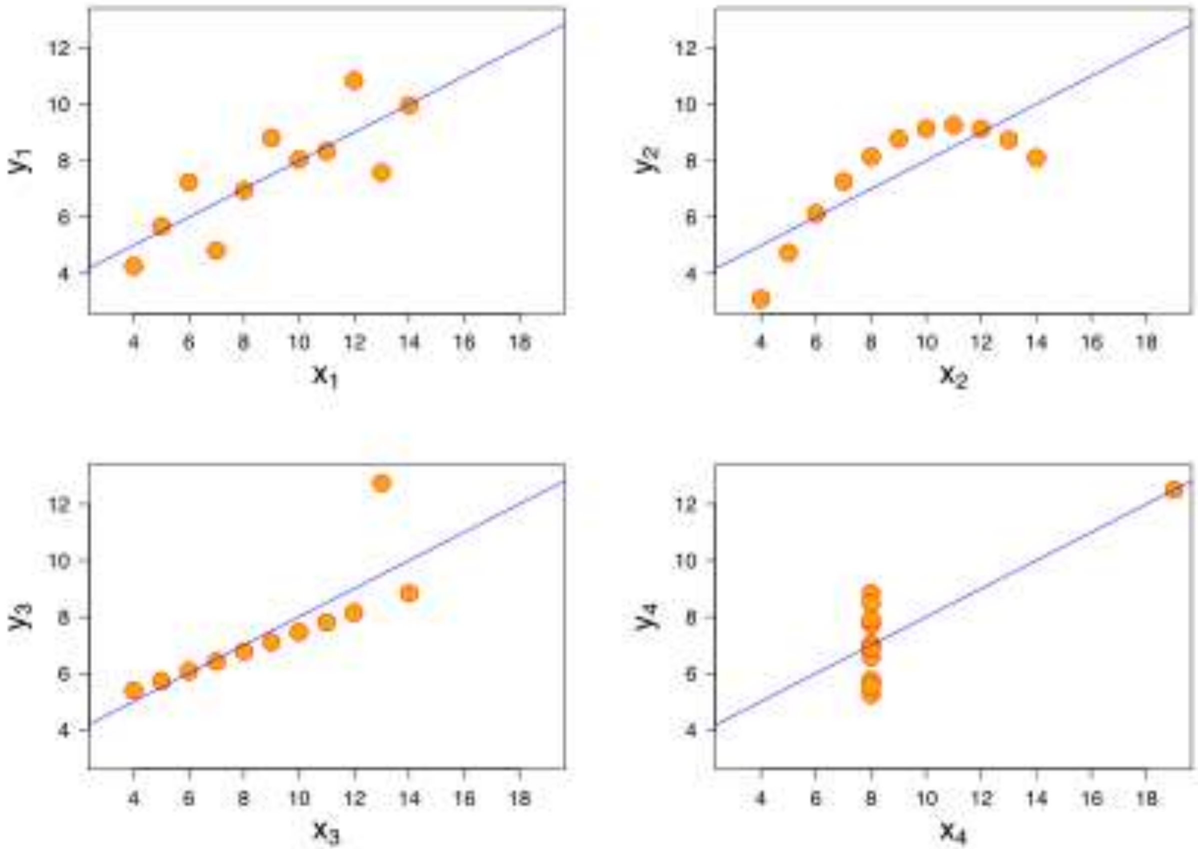
Caution with correlation: Anscombe's quartet

- The correlation coefficient can be a misleading value, and graphical examination of the data is important.

- This was illustrated beautifully by a British statistician named Frank Anscombe – by showing four examples that graphically appear very different – even though they produce identical correlation coefficients.

- These examples are famously called Anscombe's quartet.

Caution with correlation: Anscombe's quartet



In each of these examples, the following quantities were the same:

- Mean and variance of *x*
- Mean and variance of *y*
- Correlation coefficient *r*(*x*,*y*)

But the data are graphically very different!

Image source

Reflective (or Uncentered) correlation coefficient

- A version of the correlation coefficient in which you do not deduct the mean values from the vectors!

$$r(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i-\mu_x)^2\sum_{i=1}^{N}(y_i-\mu_y)^2}} \neq r_{uncentered}(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{N}x_i y_i}{\sqrt{\sum_{i=1}^{N}x_i^{\,2}\sum_{i=1}^{N}y_i^{\,2}}}$$

- Uncentered c.c. is not “translation invariant”:

$$r(\boldsymbol{x},\boldsymbol{y}) = r(\boldsymbol{x}+a,\boldsymbol{y}+b)$$

$$r_{uncentered}(\boldsymbol{x},\boldsymbol{y}) \neq r_{uncentered}(\boldsymbol{x}+a,\boldsymbol{y}+b)$$

Correlation does not **necessarily** imply causation

- A high correlation between two attributes does not mean that one causes the other.

- Example 1: Fast rotating windmills are observed when the wind speed is high. Hence can one say that the windmill rotation produces speedy wind? (a wind**mill** in the literal sense ☺)

Correlation does not **necessarily** imply causation

- In example 1, the cause and effect were swapped. High wind speed leads to fast rotation and not vice-versa.

- Example 2: High sale of ice-cream is correlated with larger occurrence of drowning. Hence can one say that ice-cream causes drowning?

- In this case, there is a third factor that is highly correlated with both – ice-cream sales, as well as drowning. Ice-cream sales and swimming activities are on the rise in the summer!

Correlation does not **necessarily** imply causation

- The above statement does not mean that correlation is *never* associated with causation (example: increase in age does cause increase in height in children or adolescents) – just that it is not *sufficient* to establish causation.

- Consider the argument: "High correlation between tobacco usage and lung cancer occurrence does **not** imply that smoking causes lung cancer."

Correlation does not **necessarily** imply causation – but it may!

- However multiple observational studies that eliminate other possible causes do lead to the conclusion that smoking causes cancer!
  - ❑ higher tobacco dosage associated with higher occurrence of cancer
  - ❑ stopping smoking associated with lower occurrence of cancer
  - ❑ higher duration of smoking associated with higher occurrence of cancer
  - ❑ unfiltered (as opposed to filtered) cigarettes associated with higher occurrence of cancer
- See https://www.sciencebasedmedicine.org/evidence-in-medicine-correlation-and-causation/ and http://www.americanscientist.org/issues/pub/what-everyone-should-know-about-statistical-correlation for more details.

More examples

Relationship between continuous and discrete variables

Future topics

- Multi-variate visualization
- Commercial systems for data visualization
- Visualizing special data
  - Time series
  - Text, e.g. point clouds