Getting to know your data

Data objects and attribute types

Basic statistical descriptions of data

The Alexander Kofkin
Faculty of Engineering
Bar-Ilan University

Based on Data Mining – Concept and Techniques,
Jiawei Han, Micheline Kamber & Jian Pei

Data objects and attribute types

The Alexander Kofkin
Faculty of Engineering
Bar-Ilan University

What is data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
  - Numeric, categoric, etc.
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, instance, or row.

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Target variable

Transaction data

- A special type of record data, where
    - each record (transaction) involves a set of items.
    - for example, set of products purchased by a customer

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Data quality

- Coverage – rows
- Completeness – columns
- Cleanliness – error in your data
- Timeliness – are the data up-to-date
- Consistency – are the data presented in the same format?

Types of attributes

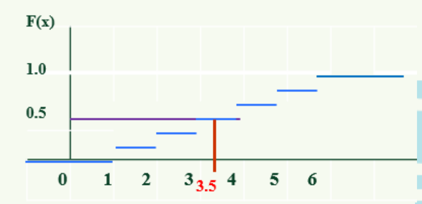
- **Nominal**
  - Examples: ID numbers, eye color, zip codes
- **Ordinal**
  - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), height in {tall, medium, short}
- **Interval**
  - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- **Ratio**
  - Examples: length, counts

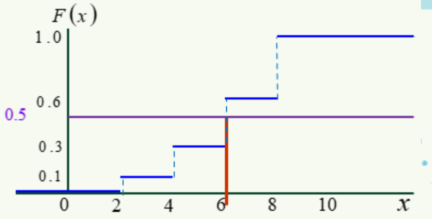
Basic statistical description of data

The Alexander Kofkin
Faculty of Engineering
Bar-Ilan University

Measuring the central tendency

- **<u>Mean (algebraic measure) (sample vs. population):</u>**
  - Note: *n* is sample size and *N* is population size: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \mu = \frac{\sum x}{N}$
  - Weighted arithmetic mean: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
- **<u>Median</u>:**
  - $\mathrm{Med(X)}=x$ such that $\mathrm{P}\{\mathrm{X} \le x\} \ge 0.5$ and $\mathrm{P}\{\mathrm{X} \ge x\} \ge 0.5$
  - Middle value if odd number of values, or average of the middle two values otherwise:
  - Estimated by interpolation (for *grouped data*):





$$median = L_1 + \left(\frac{n/2 - \sum_{i=1}^{l} f_i}{f_m}\right) c$$

n – # of entities ; L_1 – *lower limit of median interval*

f_i – frequency within interval i;

f_m – frequency within the median interval m;

c – median interval length

- **<u>Mode:</u>**
  - Value that occurs most frequently in the data
  - Empirical formula: $mean - mode = 3 \times (mean - median)$

Example
median based on grouped data

Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows:

Age	Frequency
1-5	200
5-15	450
15-20	300
20-50	1500
50-80	700
80-110	44

Compute an *approximate median* value for the data.

Answer:

Using Equation $median = L_1 + (\frac{n/2 - (\sum f)l}{f_{median}})c$, we have $L_1 = 20$, $N = 3194$, $(\text{Sum}(freq))_l = 950$, $freq_{median} = 1500$, $width = 30$, $median = 32.94$ years. 20+((1597-950)/1500)*30.

Discrete distributions

- X~U(N)

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

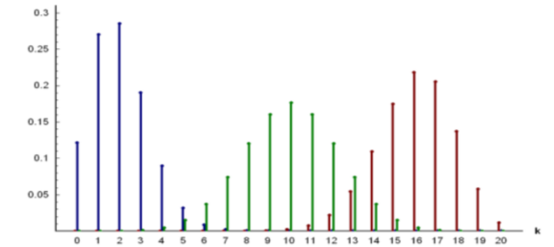
$Mode$ אין

$$V[X] = \frac{N^2-1}{12}$$ $$E[X] = \frac{N+1}{2}$$

- X~Bin(n,p)

$$\lfloor np \rfloor \le \text{med} \le \lceil np \rceil$$ $$\text{mode} = \lfloor (n+1)p \rfloor$$

$$V[X] = np(1\text{-}p)$$ $$E[X] = np$$



Bin(20,0.1)

Bin(20,0.5)

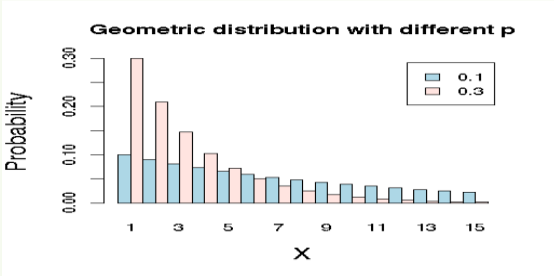
Bin(20,0.8)

- X~G(p)

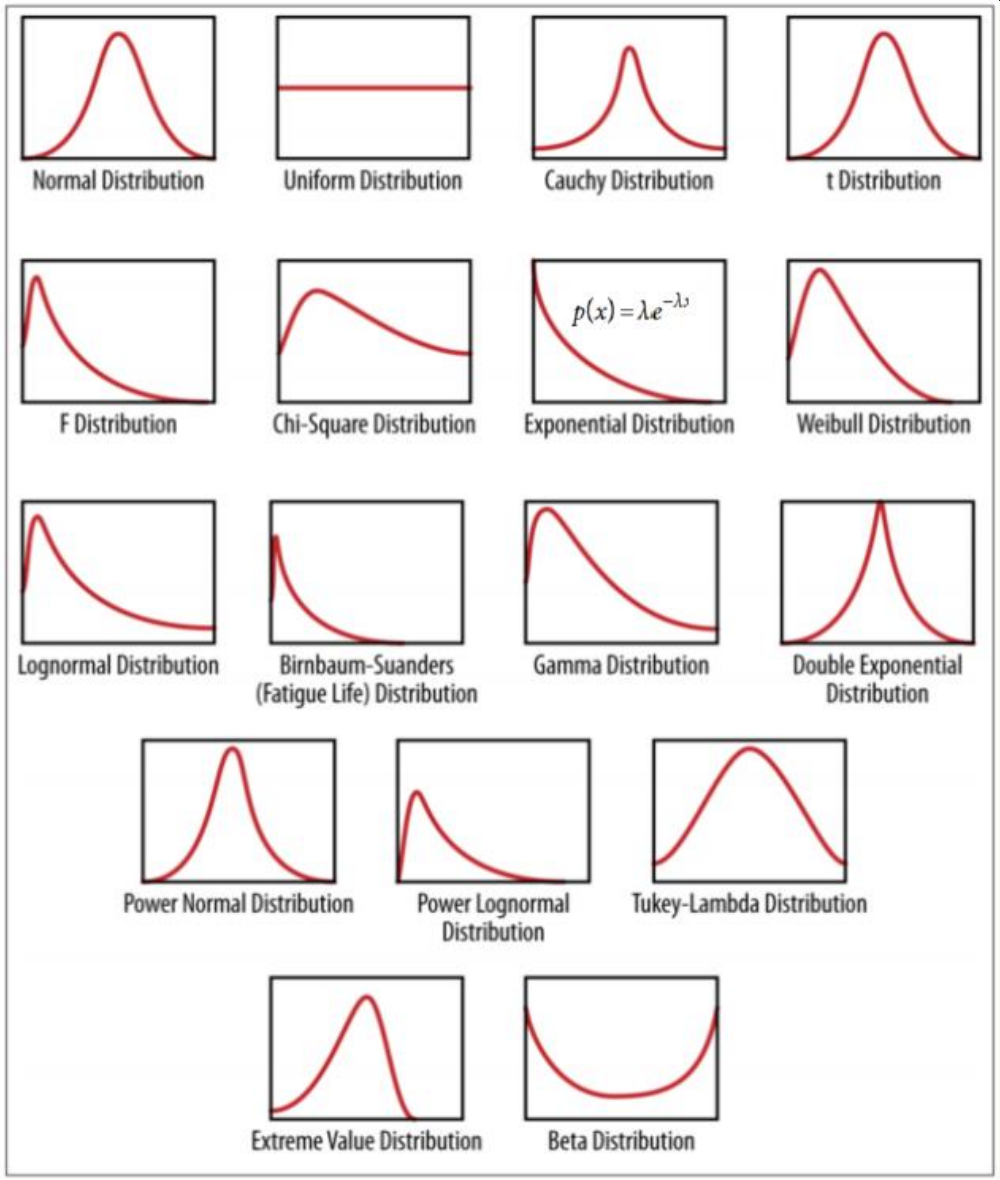
$$\frac{-\ln(2)}{\ln(1-p)}$$

Mode 1

$$V[X] = \frac{1-p}{p^2}$$ $$E[X] = \frac{1}{p}$$



Continuous distributions



Symmetric vs. Skewed data

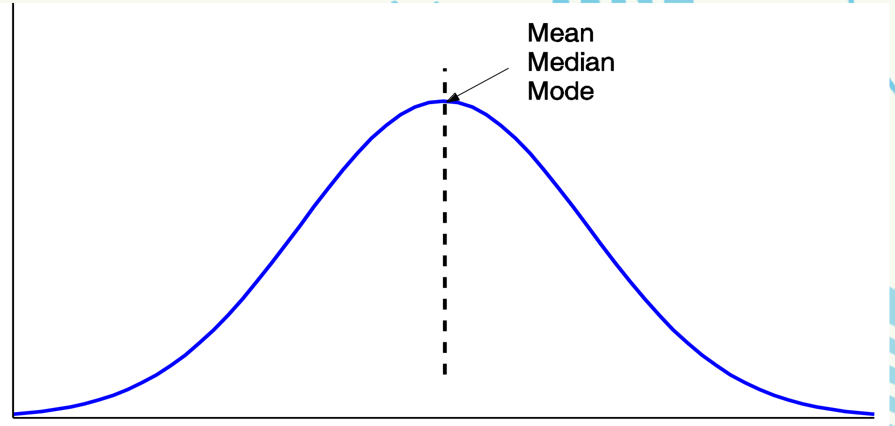
Median, mean and mode of symmetric, positively and negatively skewed data

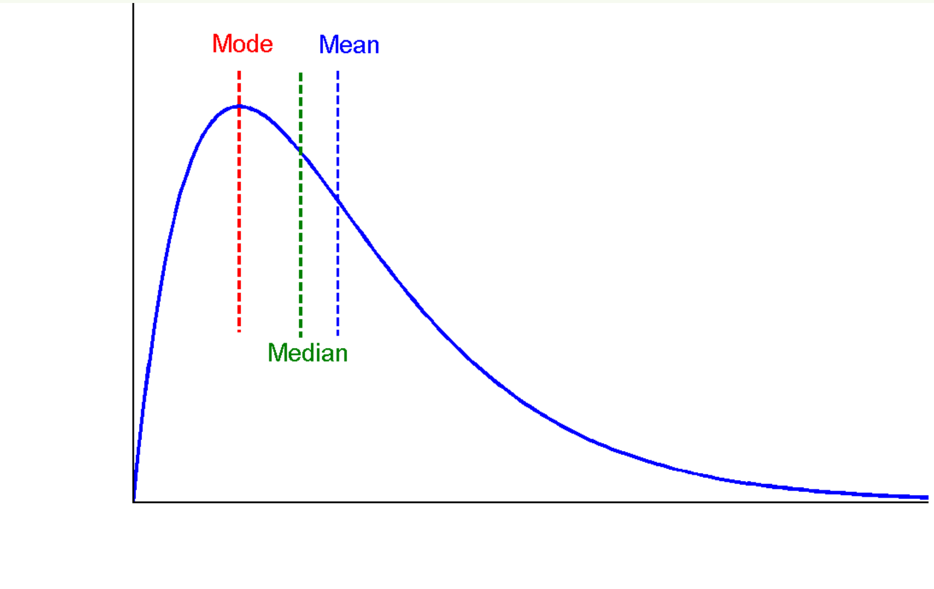
X - The predictor variable

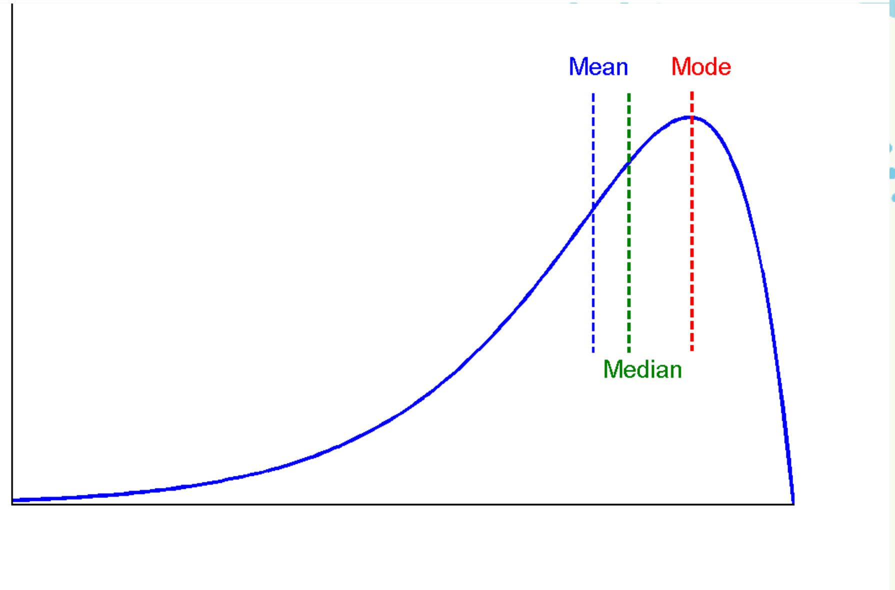
n - The # of variables

$$\text{skewness} = \frac{\sum (x_i - \overline{x})^3}{(n-1)v^{3/2}}$$

$$\text{where} \quad v = \frac{\sum (x_i - \overline{x})^2}{(n-1)}$$



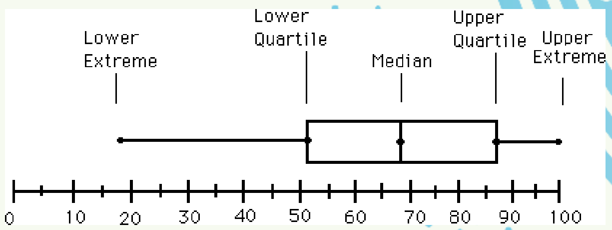


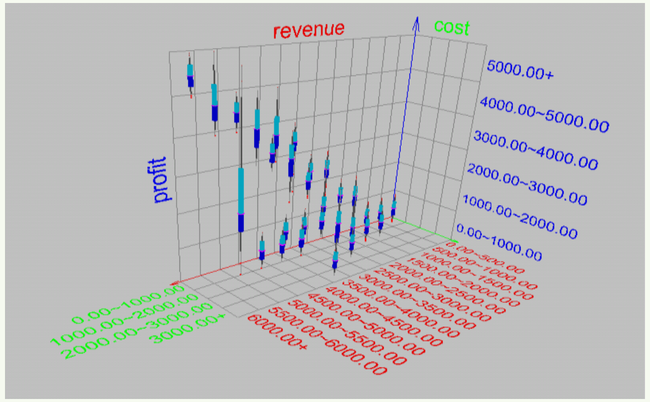


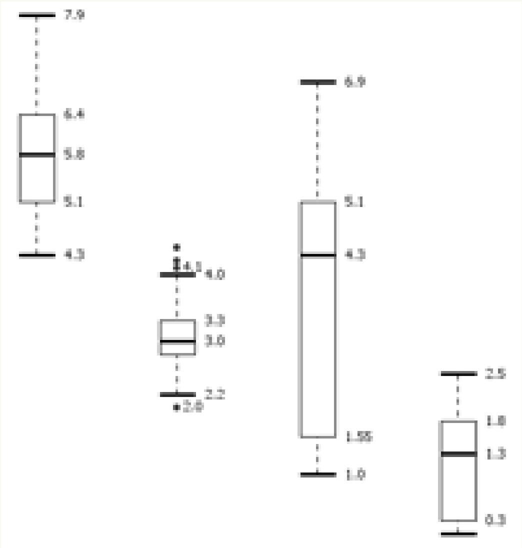
Replace the data with the log, square root or inverse may help to remove the skew.

Dispersion of data

- **Quartiles, outliers and boxplots**
  - Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
  - Inter-quartile range: IQR = $Q_3 – Q_1$
  - Five number summary: min, Q_1, median, Q_3, max
  - Boxplot: ends of the box are the quartiles; median is marked;
  - Outlier: usually, a value higher/lower than 1.5 x IQR







Dispersion of data

- **Variance and standard deviation (*sample: s, population: σ*)**

$$V[X] = E[(X-\mu_X)^2] = \sum_k (k-\mu_X)^2 \cdot P\{X=k\}$$

$$\sigma_X = \sqrt{V[X]}$$

a. $V[X]=E[X^2]-(E[X])^2$
b. $V[X]\geq 0$
c. $V[aX+b]=a^2V[X]$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^{\,2} - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right]$$

Central Limit Theorem

- N i.i.d. random variables X_i with mean μ, variance σ^2
- Assume : $\bar{x}_N = \frac{1}{N}\sum_i x_i$

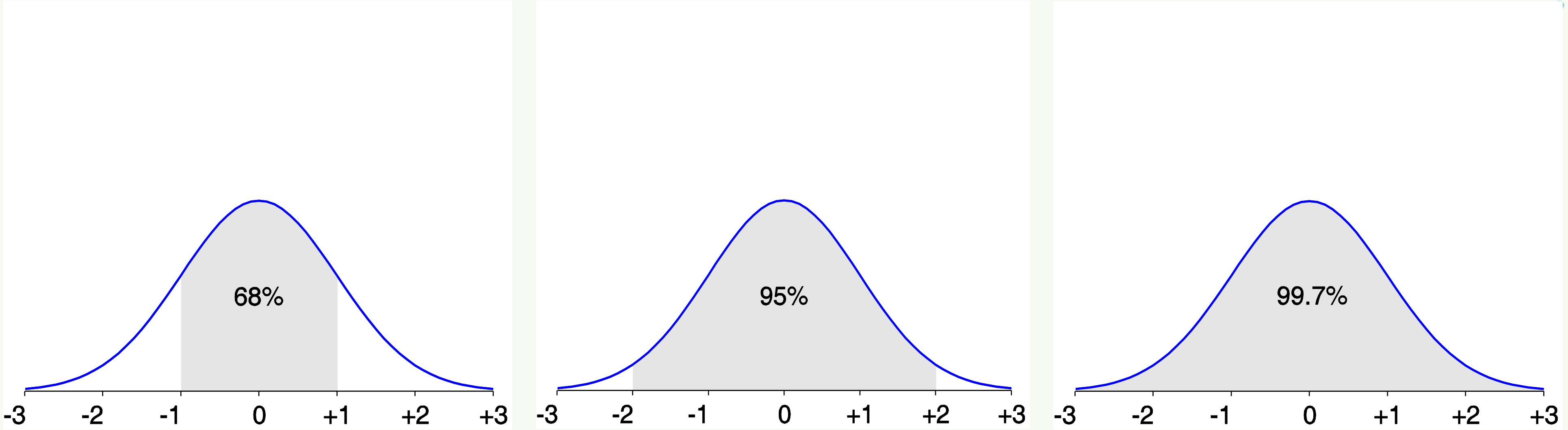
  a. $E(\bar{x}_N) = \mu$

  b. $Var(\bar{x}_N) = \frac{1}{N}\sigma^2$

- According to the central limit theorem : $f(\bar{x}_N) \sim \mathcal{N}\left(\mu, \frac{1}{N}\sigma^2\right)$
- Assume $Z_N = \frac{\bar{x}_N - \mu}{\sigma/\sqrt{N}}$ then, $f(Z_N) \sim \mathcal{N}(0,1)$

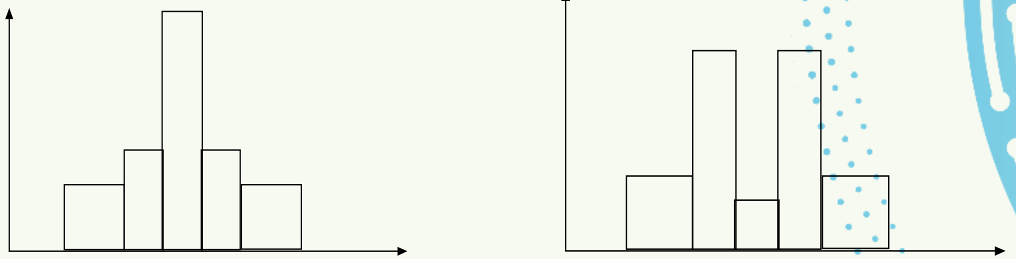
Properties of normal distribution curve

- **The normal (distribution) curve**
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it

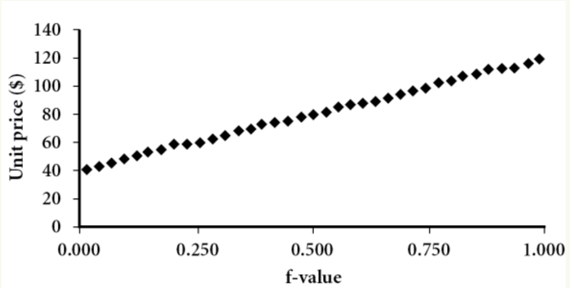


Graphic Displays of Basic Statistical Descriptions

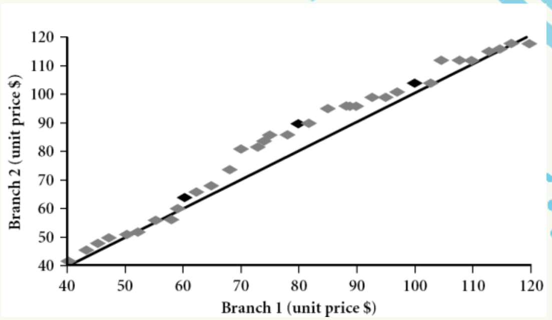
- **Boxplot**
- **Histogram**
- **Quantile plot**
- **Quantile-quantile (q-q) plot**
- **Scatter plot**



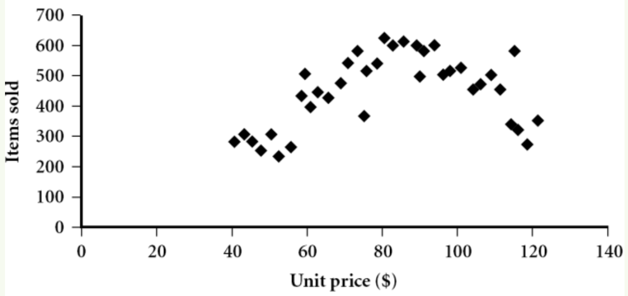
Histograms: The same boxplot representation (min, Q1, median, Q3, max)



Quantile plot: % of the data are below or equal to the value (increasing order)



q-q plot: % quantiles of one univariate distribution against the corresponding quantiles of another



Scatter plot: Each pair of values is treated as a pair of coordinates

Covariance and Correlation co-efficient

Covariance :

$$Cov(x,y) = E\{(x-\mu_x)(y-\mu_y)\} = E(x\cdot y) - \mu_x\mu_y$$

a. if x and y are independent then : $E(x\cdot y) = E(x)E(y) = \mu_x\mu_y$ → Cov(x,y) = 0

Correlation co-efficient

$$\rho(x,y) = \frac{Cov(x,y)}{\sigma_x\sigma_y}$$

a. if x and y are independent then : $\rho(x,y) = 0$

b. $|\rho(x,y)| \leq 1$

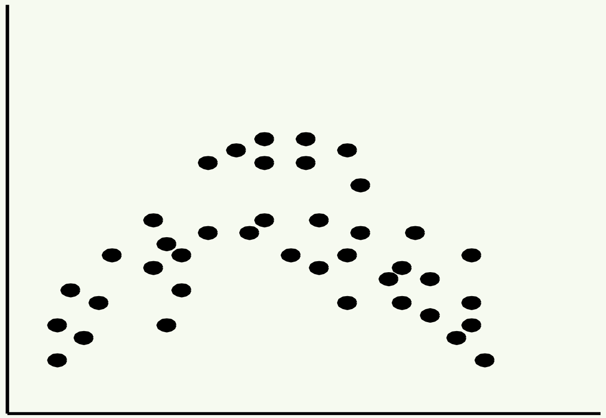
c. $|\rho(x,y)| = 1 \Leftrightarrow y = ax + b$

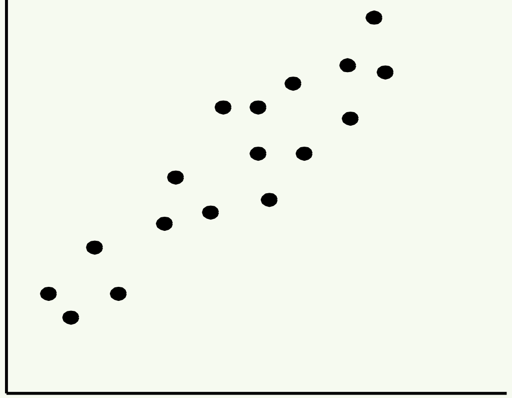
Scatter plot and correlation co-efficient

- For independent variables :

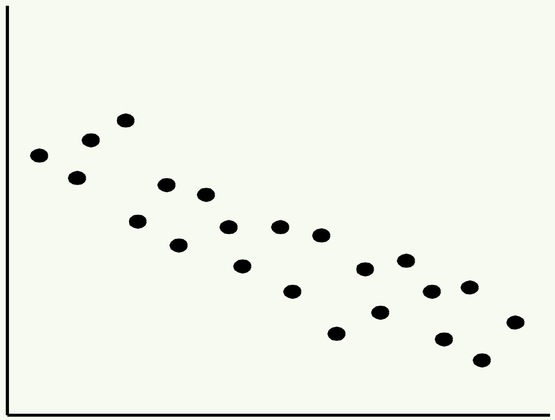
$$E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$$

- If COV(X,Y)=0 ≠>X,Y are independent ($Y = X^2$)





positively correlated



negative correlated

Thank you!

The Alexander Kofkin
Faculty of Engineering
Bar-Ilan University