CBD2204: week 4

Takis Zourntos, PhD PEng

data frames & random variables

consider a data frame, consisting of rows and labeled columns:

index	random variable 1	random variable 2	random variable 3	random variable 4	…
sample 1					…
sample 2					…
sample 3					…
⋮	⋮	⋮	⋮		

Big Data just means that we have many samples (rows)

example (data frame)

customer data

customer (sample index)	number of orders	total sales	gender
0001	5	238.77	M
0002	3	36.49	F
0003	8	313.28	U
0004	2	15.12	M
0005	9	1043.86	M
0006	4	422.27	F
0007	3	163.44	F
⋮	⋮	⋮	⋮

columns $\Leftrightarrow$ random variables

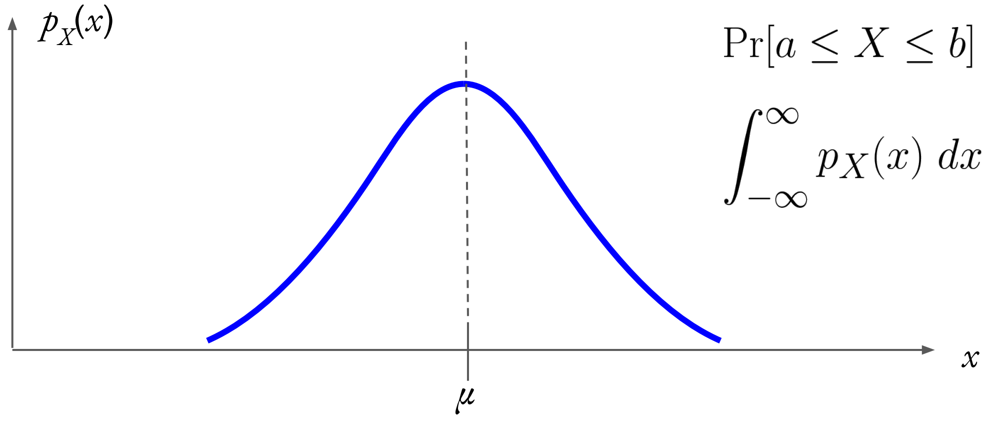
random variables

- a random variable is a mathematical concept; it refers to a sampled quantity, whose value (when sampled) cannot be predicted, but behaves according to a *probability distribution*

- think of the probability distribution as you would a histogram, since both will have the same shape, after an infinite number of samples

- for example, for any given customer (a sample), we don't know how much that customer will spend (sales) or how many times that customer will make a purchase (number of orders), but we can characterize the sales and number of orders as random variables, i.e., *statistically*.

- **example**: here is the pdf of a random variable, *X*:



$$\Pr[a \leq X \leq b] = \int_a^b p_X(x)dx$$

$$\int_{-\infty}^{\infty} p_X(x)\, dx = 1$$

*we assume that x (i.e., the value of the associated random variable), is a real quantity

random variables (expectation)

- we define the **expected value** of a random variable X in terms of its pdf as:

$$E[X] \;=\; \int_{\forall x} x\, p_X(x)\, dx$$

- which captures the intuition of a "weighted average" or mean value of the random variable, it is often represented by the Greek letter ***μ***, and also referred to as the *first moment* of the pdf.
- the deviation of the pdf from the mean is described by the **variance**, defined as:

$$\mathrm{Var}[\mathrm{X}] = E[(X-\mu)^2]$$

- and gives a measure of the "spread" in values about the mean; it also called the *second moment*; the square-root of the variance is called the **standard deviation**:

$$\sigma = \sqrt{E[(X-\mu)^2]}$$

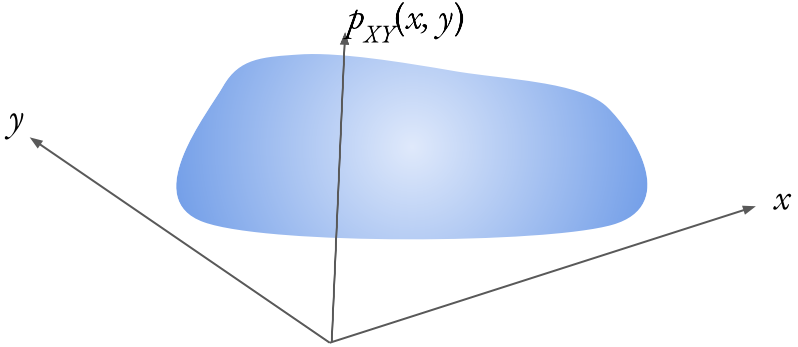
- as a consequence, the variance is sometimes represented by σ^2

random variables (cont'd)

- *dependence*: although they are not predictable, two random variables can have probability distributions which *depend* on each other
- two random variables, X and Y, are ***independent*** iff

$$p_{XY}(x, y) = p_X(x)\, p_Y(y), \;\; \forall\; x,\; y$$

- in which $p_X()$ is pdf of X, $p_Y()$ is the pdf of Y, and $p_{XY}()$ is the joint probability distribution



random variables (cont’d)

- if two random variables are *independent*, then they are *uncorrelated*. However, *correlated* variables are not necessarily dependent, in the probabilistic/mathematical sense.
- we define the *covariance* of random variables X and Y as:

$$\operatorname{cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

- we can use covariance to determine the ***correlation coefficient***, between two random variables:

$$\rho_{XY} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- **which gives a measure of the quality of a least-squares fit between the variables**

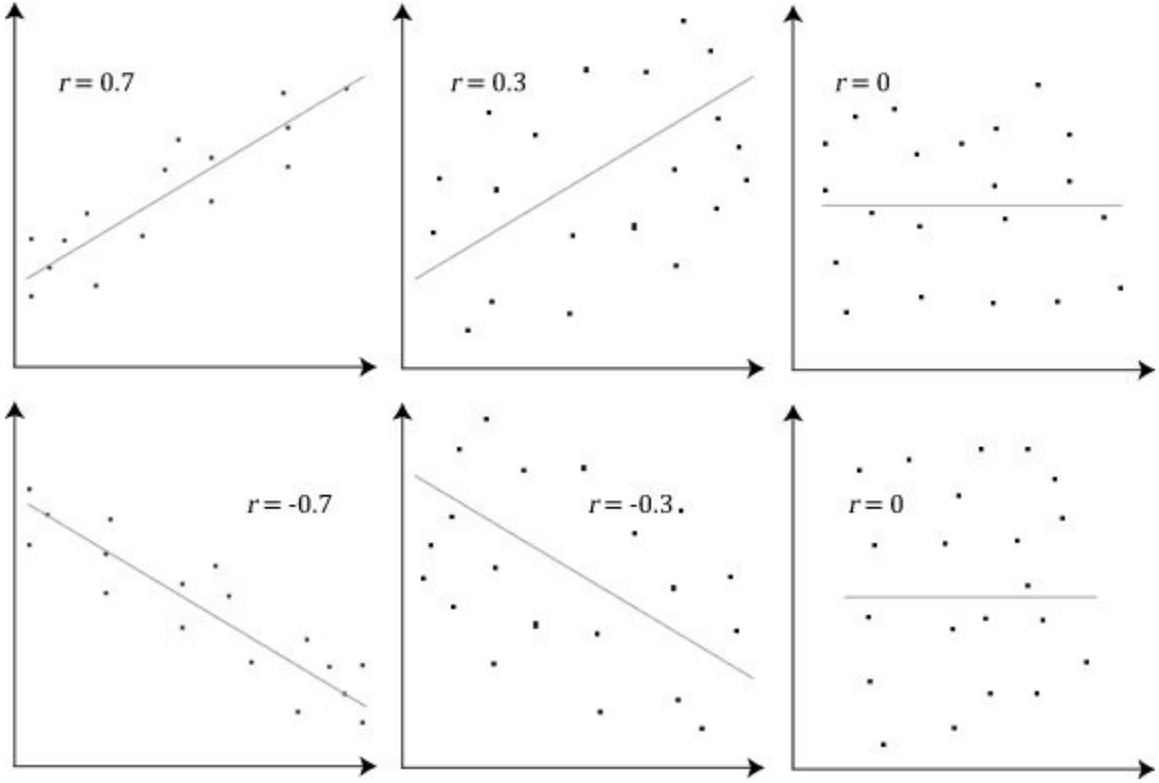
sample calculations

$$E[X] \approx \frac{1}{N}\sum_{i=1}^{N} x_i \quad =: m_X$$

sample version of correlation coefficient:

$$r = \frac{1}{N}\frac{\sum_{i=1}^{N}(x_i - m_X)(y_i - m_Y)}{s_X s_Y}$$

it helps if *N* is large!


r = 0.7
r = 0.3
r = 0
r = -0.7
r = -0.3
r = 0

descriptive statistics in *R*

recall that the *summary*() function in *R* provides several descriptive stats about the variables (columns) in a data frame:

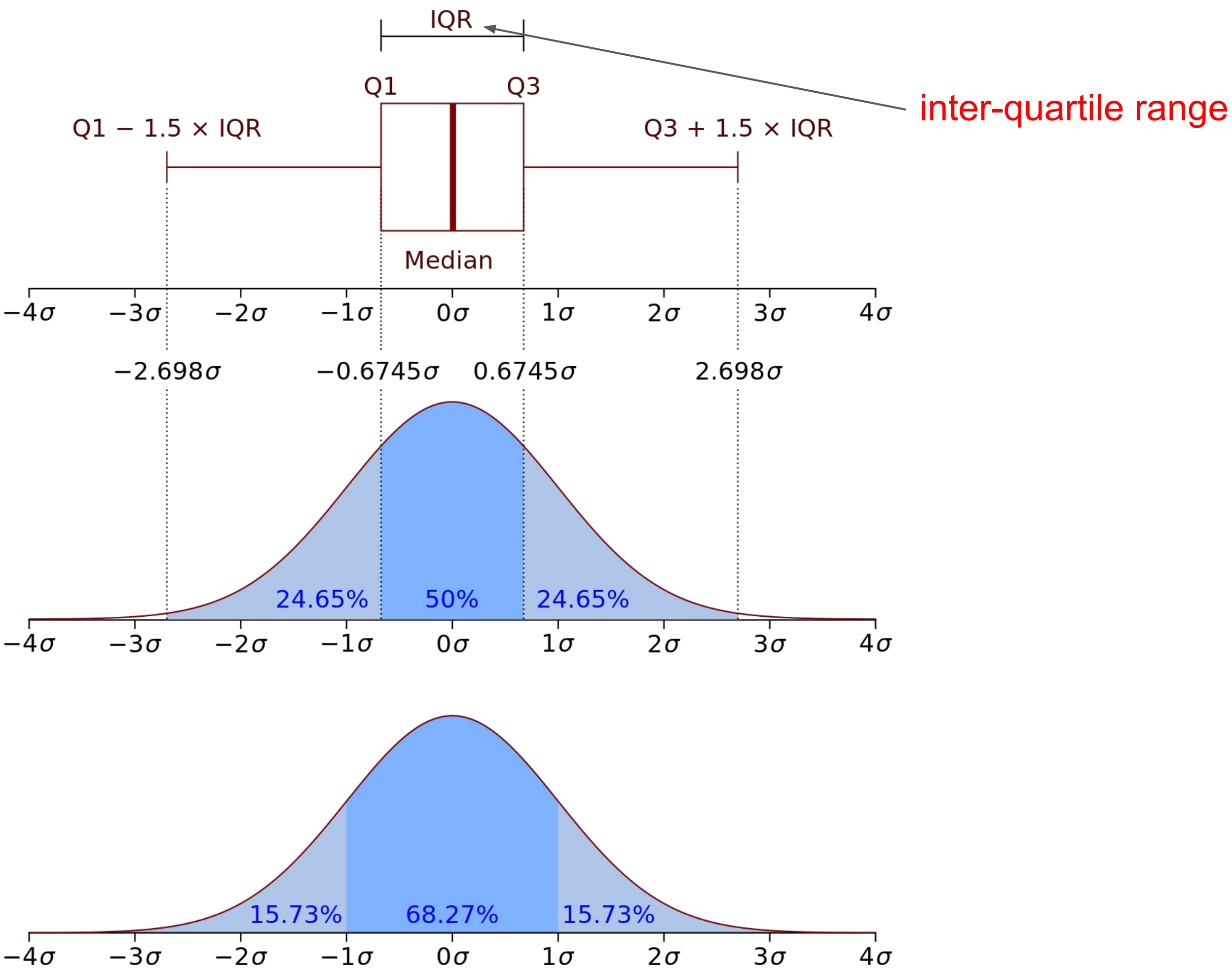
- mean
- median
- quartiles

for median and quartiles, data is *ordered*; for an odd-number of samples, **the median is the middle sample point, and for an even number of samples, the median is the average of the smallest and largest samples**.

- example: consider our *number of sales* data above: {5, 3, 8, 2, 9, 4, 3}
- ordering yields:
    - {2, 3, 3, 4, 5, 8, 9}

median ($q_2=4$)

$q_1=3$ $q_3=6.5$



https://commons.wikimedia.org/wiki/File:Boxplot_vs_PDF.svg

back to *R*

check out the following functions in *R*:

- summary()
- mean()
- median()
- cov()
- cor()