Probability and Statistics Review

2019. 6

Reference:

- Steven Skiena, The Data Science Design Manual, Springer, 2017
- https://www.cs.cmu.edu/~epxing/Class…/Probability_Review.ppt

Probability and Statistics

- Probability
  - deals with predicting the likelihood of future events
  - theoretical branch of mathematics on the consequences of definitions
  - For dice game, “each face will come up with probability 1/6.”
- Statistics
  - analyzes the frequency of past events
  - applied mathematics trying to make sense of real-world observations
  - For dice game, “I will watch a while, and keep track of how often each number comes up.”

Probability

- Experiment: a procedure which yields one of a set of possible outcomes
- Sample space S: set of possible outcomes *s* of an experiment
- Event: specified subset of the outcomes of an experiment
- Probability *p(s)* of an outcome *s: a number with*:
  - *0 <= p(s) <= 1*
  - $\sum_{s \in S} p(s) = 1$
- Random variable V: numerical function(assignment) on the outcomes of a probability space
- Expected value E of a random variable V on sample space S:

$$E(V) = \sum_{s \in S} p(s) \cdot V(s)$$

Probability (example)

- Experiment: tossing two six-sided dice
- Sample space S: 36 possible outcomes, namely
  - S = {(1,1),(1,2),(1,3),(1,4),(1,5),(1,6), (2,1),(2,2),(2,3),(2,4),(2,5),(2,6), (3,1),(3,2),(3,3),(3,4),(3,5),(3,6), (4,1),(4,2),(4,3),(4,4),(4,5),(4,6), (5,1),(5,2),(5,3),(5,4),(5,5),(5,6), (6,1),(6,2),(6,3),(6,4),(6,5),(6,6)}
- Event: the event that the <u>sum of the two dice</u> equals 7 or 11
  - E = {(1,6),(2,5),(3,4),(4,3),(5,2),(6,1), (5,6),(6,5)}
- Probability of the event: *p(E)* = 8/36
- Random variable V: V(s) = 2 , 3 , …, 12 for each sample *s*
  - P(V=7) = 6/36, p(V=12) = 1/36, p(V=7 or V=11) = 2/9
- Expected value E:
  - E(V) = 1/36(2) + 2/36(3) + 3/36(4) +4/36(5) + 5/36(6) + 6/36(7) + 5/36(8) + 4/36(9) + 3/36(10) + 2/36(11) + 1/36(12)

Compound Events and Independence

- Suppose half my students are female (event A), and Half my students are above median (event B). What is the probability a student is both A & B?
- Events A and B are independent iff

$$P(A \cap B) = P(A) \times P(B)$$

- Independence (zero correlation) is good to simplify calculations but bad for prediction (no information shared between events *A* and *B*)

Conditional Probability

- The conditional probability P(A|B) is defined:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

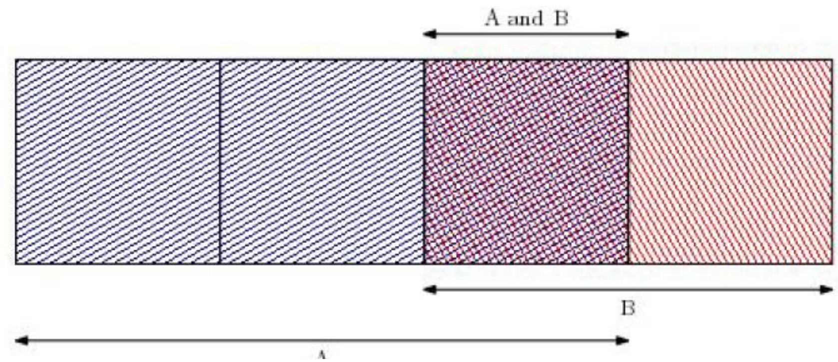
- Conditional probability get interesting only when events are ***not*** independent, otherwise:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

Bayes Theorem

- Bayes theorem is an important tool which reverses the direction of the dependences:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$



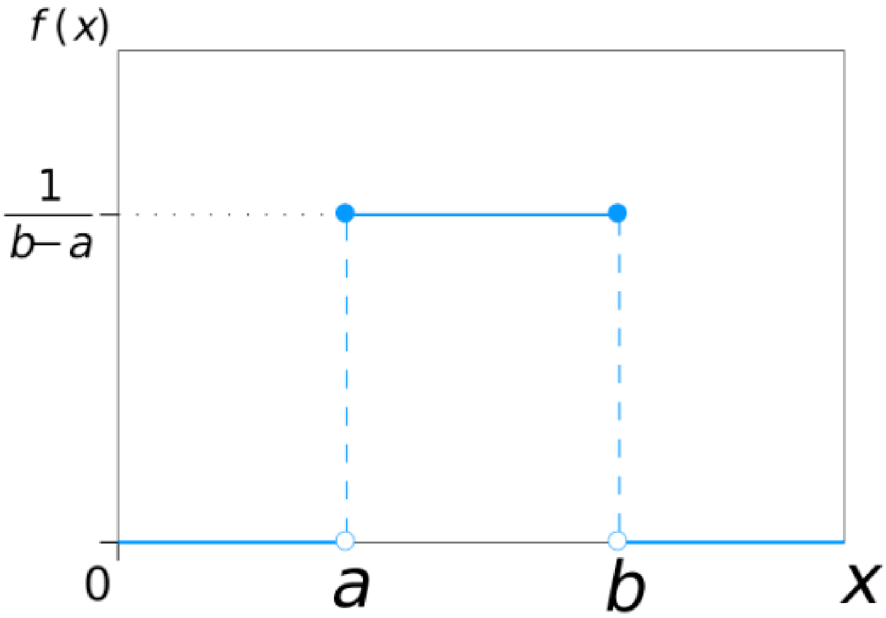
Probability Distribution

- 확률분포
  - 확률변수 x 가 특정한 값을 가질 확률 정보
- Probability Density Function, PDF (확률밀도함수)
  - 연속 확률변수에서 확률변수의 분포
  - (ex) 키, 나이
- Probability Mass Function, PMF (확률질량함수)
  - 이산확률변수에서 특정값에 대한 확률
  - (ex) 주사위, 동전
- Cumulative Distribution Function, CDF (누적분포함수)

$F_X(x) = P(X \le x)$ $\quad$ $F_{X,Y}(x, y) = P(X \le x, Y \le y)$

Probability Distribution

- Uniform Distribution(균일분포)



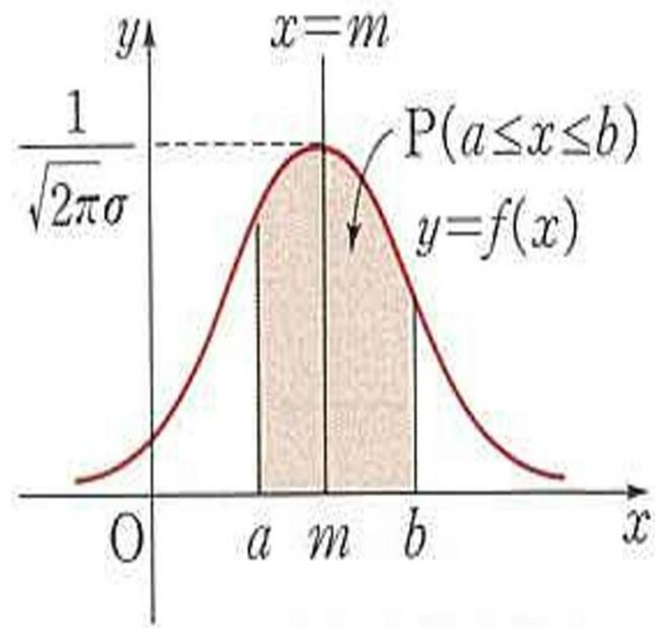
Probability Distribution

- Normal Distribution (정규분포)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} \quad (x\text{는 모든 실수})$$

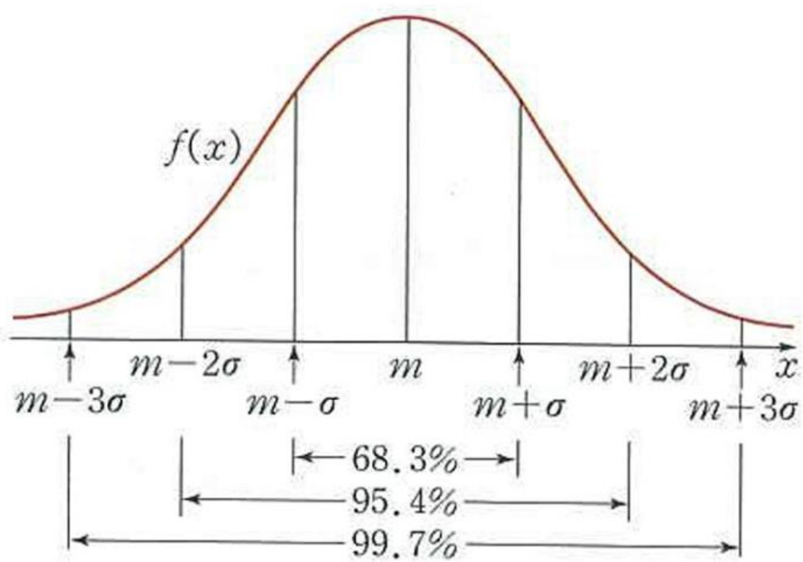
  – 확률 P(a ≤ X ≤ b) =

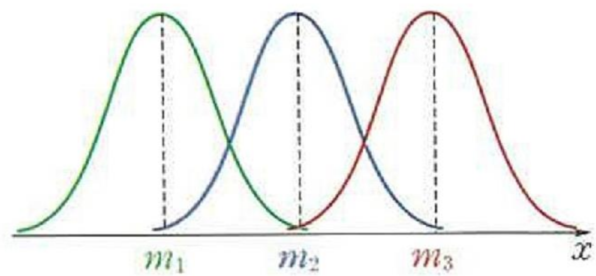
$$\int_a^b f(x)dx = \int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} dx$$

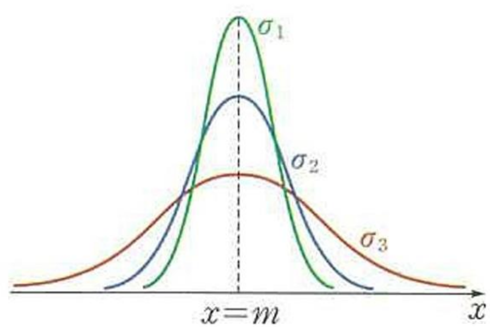


Probability Distribution

- Normal Distribution (continued)





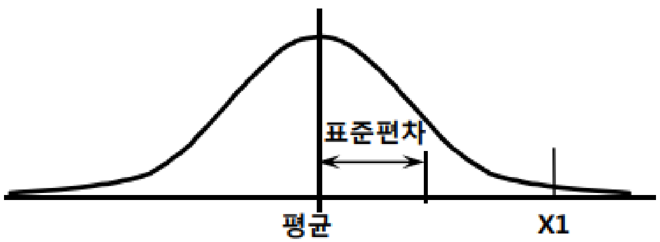


Probability Distribution

- 표준정규분포(Standard Normal Distribution)

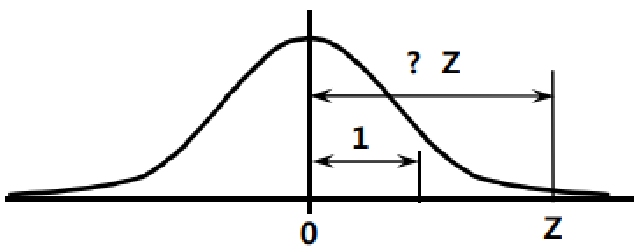
  – **정규분포(평균 μ, 분산σ²)**

  확률변수 X는 $X \sim N(\mu, \sigma^2)$



  – **표준정규분포(평균0, 표준편차1)**

  확률변수 Z은 $Z \sim N(0,1)$



정규분포

$$X \sim N(\mu, \sigma^2)$$

$$Z_i = \frac{x_i - \mu}{\sigma}$$

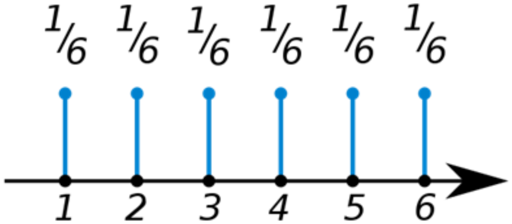
Z 변환

표준정규분포

$$Z \sim N(0, 1^2)$$

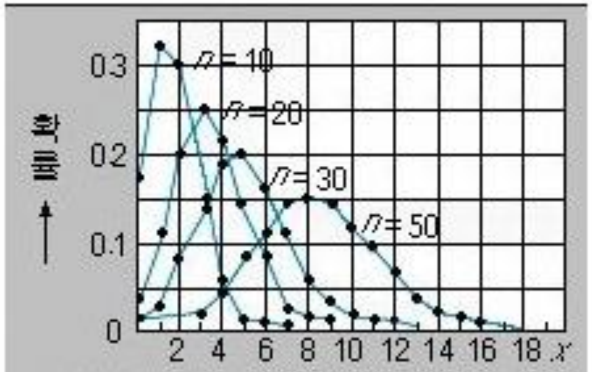
Probability Distribution

- Discrete random variable
  - (ex) 주사위를 한 번 던져 나올 값의 확률변수: X



  - 이항분포(Binomial Distribution): 여러 번의 연속 실험의 확률 (ex: 축구선수의 패널티킥 성공 확률이 0.8 일 때 10번 차서 7번 성공할 확률)

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$



1회 주사위를 10회, 20회, 30회, 50회 던졌을 때, 그 눈이 x회 나올 확률을 그래프로 나타낸 것

(*) p가 0이나 1에 가깝지 않고 n이 충분히 크면 이항분포는 정규분포(가우스분포)에 가까워지며, p가 1/2에 가까워짐에 따라 그래프는 좌우대칭인 산 모양 곡선이 된다.

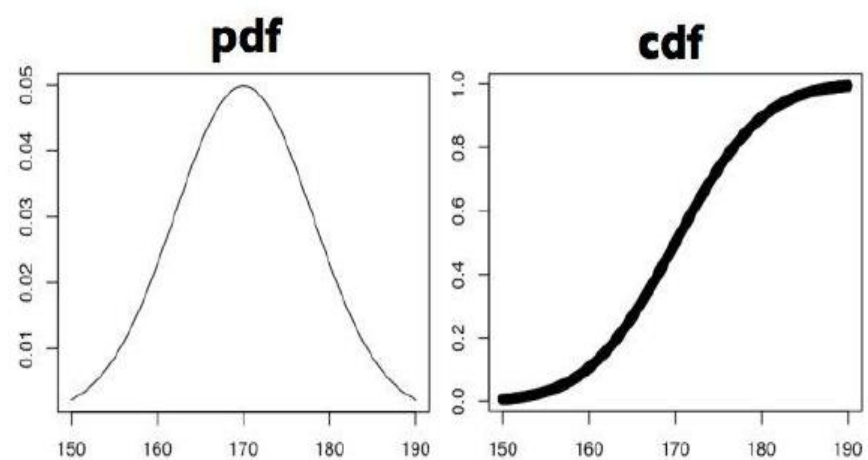
Probability/cumulative distribution

- The cdf is the running sum of the pdf:

$$F_X(x) = \mathrm{P}(X \leq x)$$

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$$

- The pdf and cdf contain exactly the same information, one being the integral / derivative of the other.

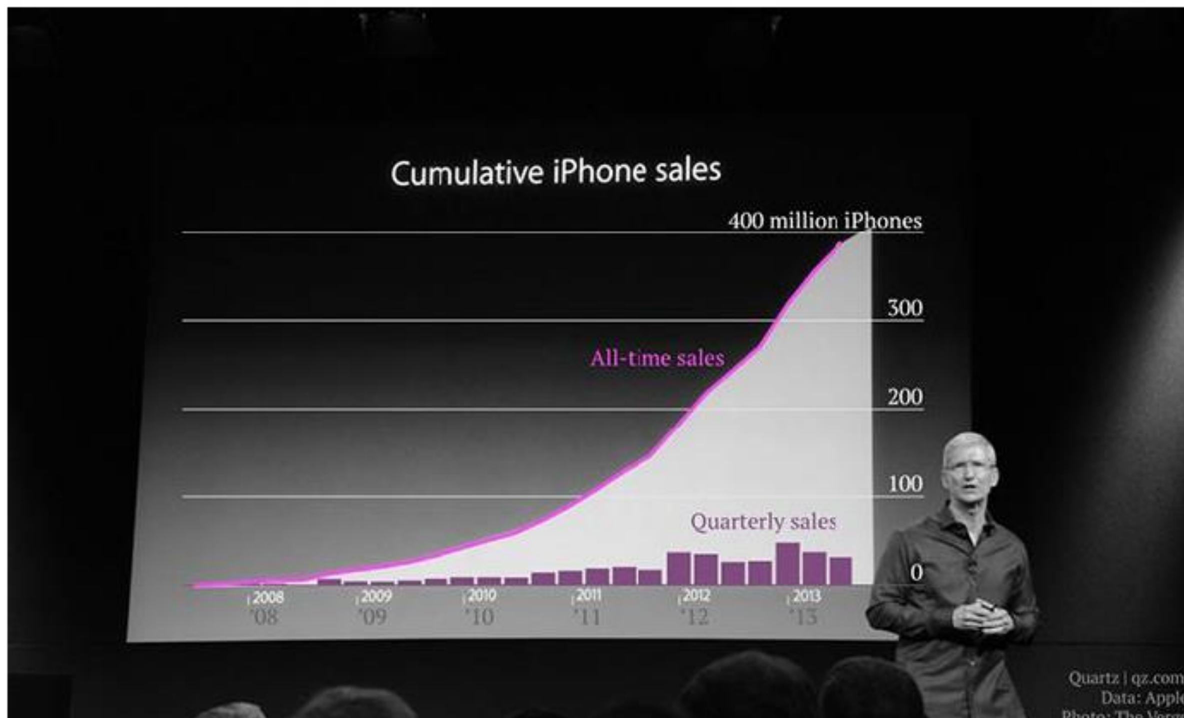


Visualizing Cumulative Distributions

- Apple iPhone sales have been exploding, right?

How explosive is that growth, really?

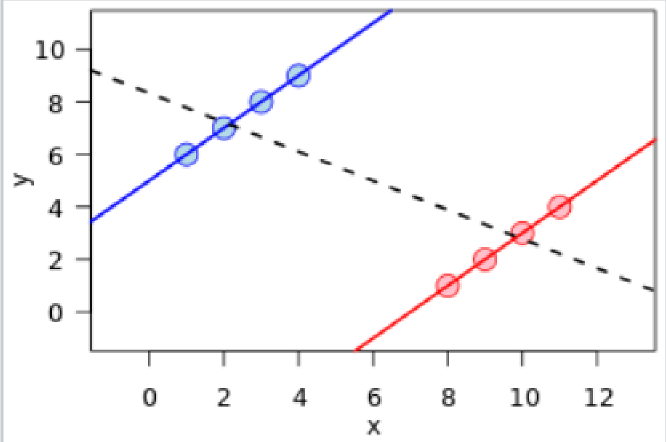
- Cumulative distributions present a misleading view of growth rate.
  - The incremental change is the derivative of this function, which is hard to visualize



Simpson's paradox

- What is it?
  - A trend appears in several different groups of data, but disappears or reverses when these groups are combined
  - Need be careful when analyzing numbers only

city	Maker A	Maker B
Seoul	Good 90 Defective 10 (rate: 10%)	Good 920 Defective 80 (rate: 8%)
Jeju	Good 980 Defective 20 (rate: 2%)	Good 99 Defective 1 (rate: 1%)
overall	defective rate: 30/1,100 = 3%	defective rate: 81/1,100 = 8%



Simpson's paradox for quantitative data: a positive trend (,) appears for two separate groups, whereas a negative trend (----) appears when the groups are combined.

(https://en.wikipedia.org/)

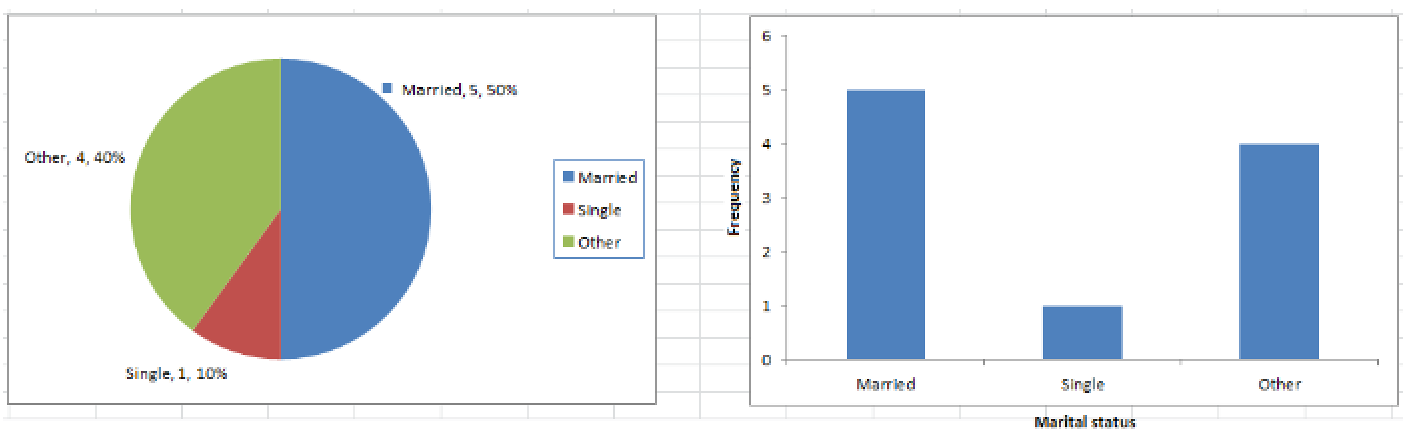
Descriptive Statistics

- Descriptive statistics provides ways to capture the properties of a given data set / sample.
  - Central tendency measures describe the center around the data is distributed.
  - Variation or variability measures describe data spread
- Centrality measure
  - Mean: arithmetic, geometric, harmonic $\mu_X = \frac{1}{n} \sum_{i=1}^{n} x_i$
  - Median
  - Mode

$$\left(\prod_{i=1}^{n} a_i \right)^{1/n} = \sqrt[n]{a_1 a_2 \cdots a_n}.$$

Uni-variate Descriptive Statistics

- Different ways you can describe patterns found in uni-variate data include
  - central tendency : mean, mode and median
  - dispersion: range, variance, maximum, minimum, quartiles , and standard deviation.
- Graphs: Pie-charts or Bar Graphs



Pie chart [left] & Bar chart [right] of Marital status from loan applicants table.

Bi-variate Descriptive Statistics

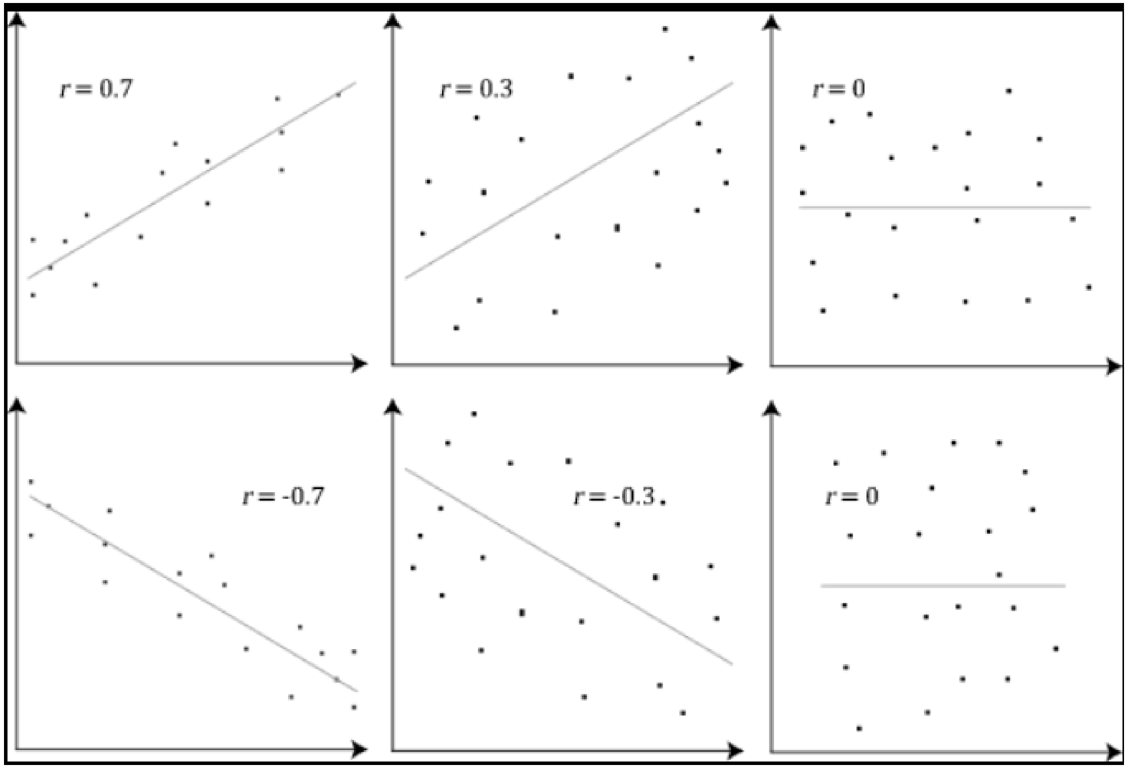
- Bi-variate analysis: analysis of two variables for the purpose of determining the empirical relationship between them.
- Graphs:
  - *Scatter Plots:* sometimes called correlation plots because they show how two variables are correlated.
  - *Bar plots*
- **Correlation**:
  - The correlation coefficient ***r*** quantifies the strength and direction of the **linear relationship** between two quantitative variables.

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)\,s_x s_y}$$

$= cov(x,y)/S_x S_y$

where `sx` and `sy` represent the standard deviation of the x-variable and the y-variable, respectively. `-1 ≤ r ≤ 1`.

Correlation



Example **scatterplots** of various datasets with various **correlation coefficients**.
(Note that correlation have no relations with the slope of the scatter plot.)
[ref: https://en.wikipedia.org/wiki/Correlation_and_dependence]

The Three Ms

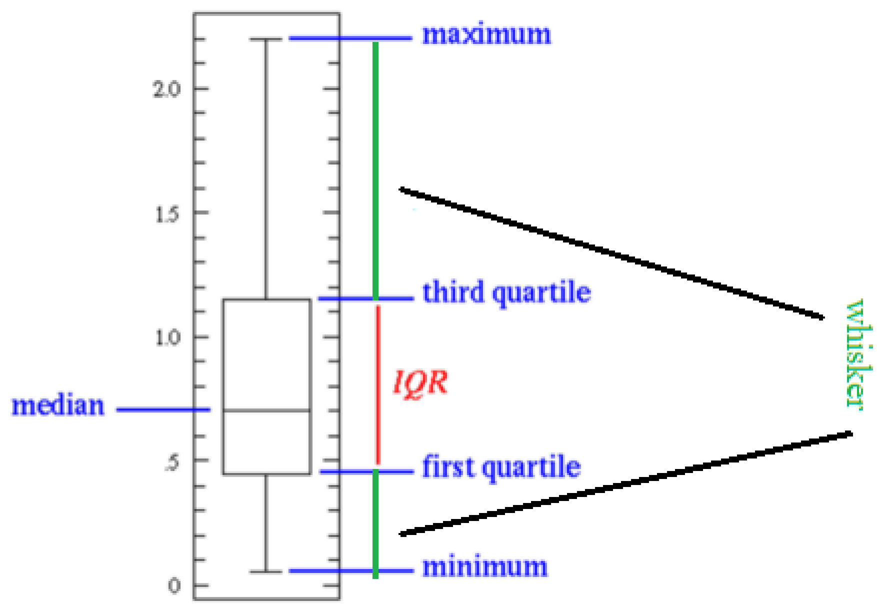
- Mean(평균): the average result
- Median(중간값): the score that divides the result in half – the middle value
- Mode(최빈치): the most common result

Which measure is best?

- Mean is meaningful for symmetric distributions without outliers: e.g. height and weight.
- Median is better for skewed distributions or data with outliers: e.g. wealth and income.
- Bill Gates adds $250 to the mean per capita wealth but nothing to the median.

Boxplots

- Box plots are especially useful for indicating whether a distribution is skewed and whether there are potential unusual observations (outliers) in the data set.
- Outliers: 1.5*IQR above the third quartile or 1.5*IQR below the first quartile

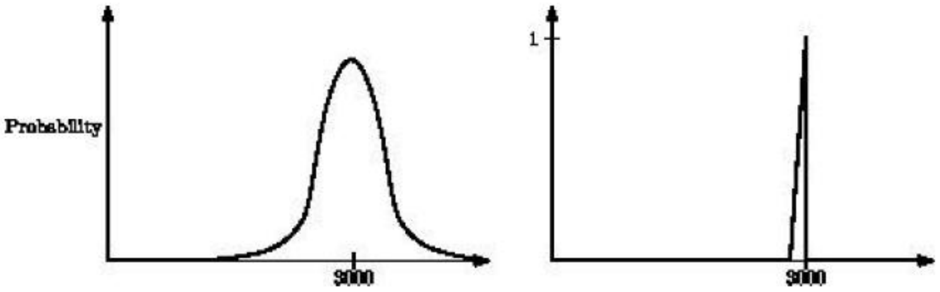


Variance Metric: Standard Deviation

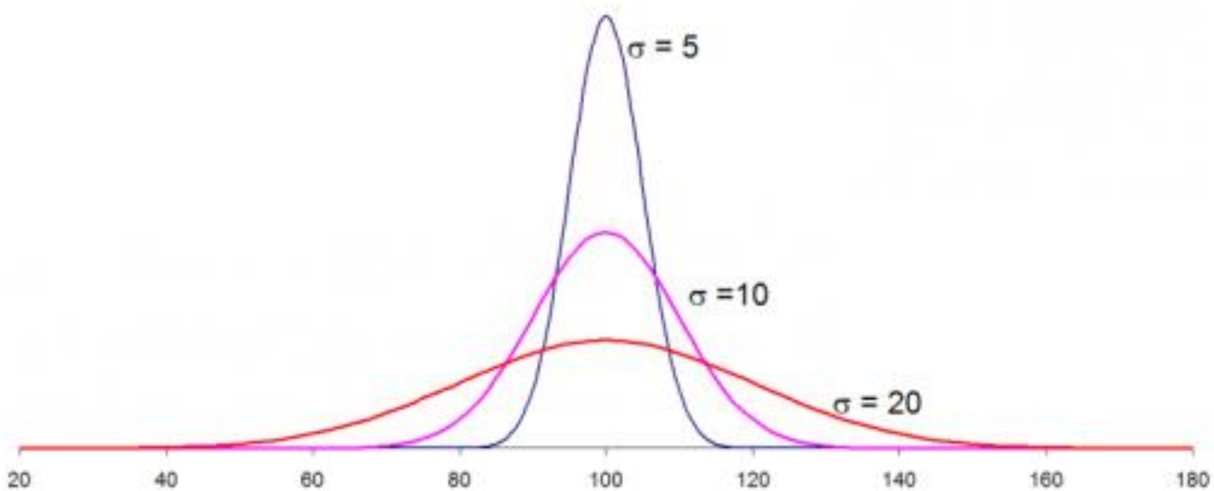
- The variance is the square of the standard deviation sigma.

$$\hat{\sigma} = \sqrt{\frac{\sum_{i}^{n}(x_i - \bar{x})^2}{n-1}}$$

- Distributions with the same mean can look very different. But together, the mean and standard deviation fairly well characterize any distribution.



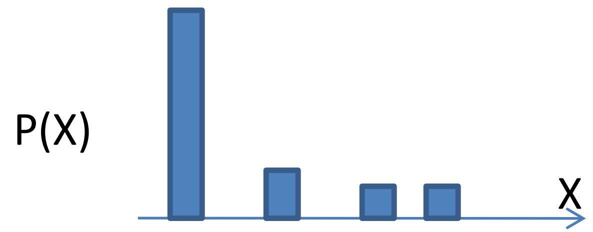
Standard Deviation



Three different data distributions with same mean (100) and different standard deviation (5,10,20)

Information Theory

- P(X) encodes our uncertainty about X
  - Some variables are more uncertain than others



  - How can we quantify this intuition?
    - Information: $\log \frac{1}{p(x)}$
    - Entropy: average number of bits required to encode X

$$H_P(X) = E\left[\log \frac{1}{p(x)}\right] = \sum_x P(x) \log \frac{1}{P(x)} = -\sum_x P(x) \log P(x)$$