BITTIGER

DS 501 Data scientist express bootcamp

Week 1 Knowledge

BITTIGER

Copyright Policy

BITTIGER

版权声明

What is statistics? Why is it important?

Statistics

- All about quantifying uncertainty (which subject is studying certainty?)
- Why uncertainty? **imperfect** data
  - Psychology
  - Geology
  - Economics
- Why imperfect data?
  - Only observe a small fraction, e.g., candidate poll
  - Only observe indirect signs, e.g., neuro signal
  - Data always contain noise, e.g, measurement error

http://www.stat.cmu.edu/~cshalizi/uADA/16/lectures/01.pdf

Probability

- Probability is central to statistics
  - Probability = Statistics?
- Basic concepts
  - **Experiment**
  - **Sample space**: all possible outcomes of an **experiment**
  - **Event:** subset of sample space

https://www.youtube.com/watch?v=KbB0FjPg0mw&list=PL2SOU6wwxB0uwwH80KTQ6ht66KWxbzTIo
from 15min

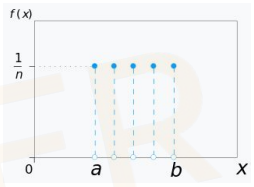
Statistics uses probability to model inference from data. We try to mathematically understand the properties of different procedures for drawing inferences: Under what conditions are they reliable? What sorts of errors do they make, and how often? What can they tell us when they work? What are signs that something has gone wrong? L

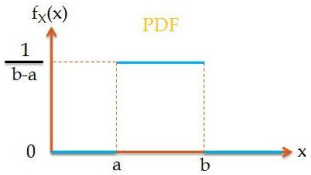
Random variable and types

- Random variable (X)
  - Enumerate sample space to real value
- Categories
  - Discrete
  - Continuous
- Examples
  - Flip a coin; Daily traffic; Number of people click on an ad; BMI
- R
  - bar chart, pie chart
  - Probability density

How to describe probability of random variable

- Probability density function (PDF)
  - PDF: x ~ P(X = x)
  - For X, frequency of possible outcome x's occurrences of the total possible occurrences
    - Example: rolling dice, P(X = 1) = ⅙
  - Discrete (pmf) v.s continuous (pdf)
  - Non-negative, sum up to 1
- Rules
  - $0 \le P(X = x) \le 1$, $P(X = \Omega) = 1$, $P(X = \varnothing) = 0$, $P(X = x) + P(X \mathrel{!}= x) = 1$, $P(X = x \text{ or } Y = y) = P(X = x) + P(Y = y) - P(X = x \text{ and } Y = y)$





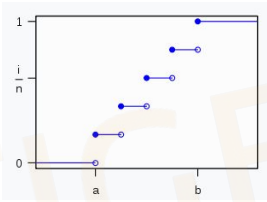
https://onlinecourses.science.psu.edu/stat414/node/97

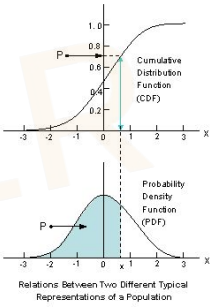
Example interview questions

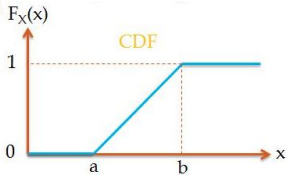
- 2 dices, probability of at least one 4 shows up
- 2 dices, probability of the sum is 6
- ...

How to describe probability of random variable(cont'd)

- Cumulative density function (CDF)
  - CDF: x ~ P(X <= x)
  - Discrete v.s continuous
  - Non-decreasing, up to 1
- Examples
  - Roll dice
  - Uniform distribution





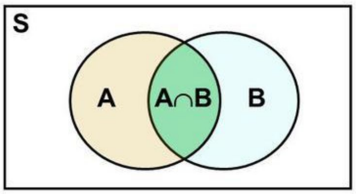


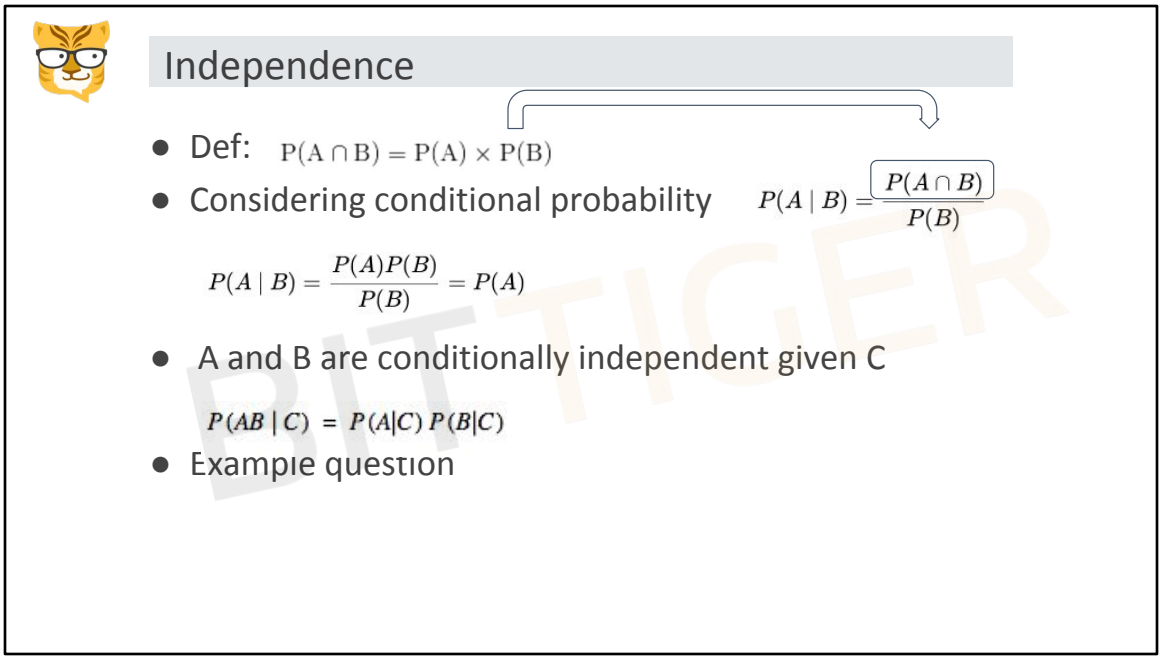
Conditional probability

- How to update uncertainty after getting new information?
  - P(dice = 1) = ⅙, what if we know it's an odd number?
- Definition, Probability of A given B

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

- Intuition
  - Only consider B
  - Renormalize
- Example question





conditional independence is the same as normal independence, but restricted to the case where you know that a certain condition is or isn't fulfilled.

Say you roll a blue die and a red die. the results of the dice rolls are conditionally independent with respect to the event "the blue result is not 6 and the red result is not 1", but they're not conditionally independent with respect to the event "the sum of the results is even".

A and B are independent <->
A and B are independent given C ?

Left to right. Independence does not imply conditional independence: for instance, independent random variables are rarely independent conditionally on their sum or on their maximum.

Right to left: A = x +C, B = y + C
P(AB) = P(A)P(B)
C = A- B, then P(AB|C) != P(A|C) P(B|C)

A = C+D, B = C+E, then P(AB|C) = P(A|C) P(B|C) but P(AB) != P(A) P(B)

Bayes' theorem

- Reverse the conditional probability

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)}$$

- Prove it
  - Hint: $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$
- Fair coin and unfair coin problem

P(fair coin | 8 heads out 10) = P(8 heads out 10 |fair coin) * P(fair coin) / [P(8 heads out 10 |fair coin) * P(fair coin) + P(8 heads out 10 | not fair coin) * P(not fair coin)]

B: Gene mutation A: medical test

http://web.ipac.caltech.edu/staff/fmasci/home/astro_refs/Science-2013-Efron.pdf

Expectation

- Measure what's the center of a random variable
  - Discrete $E[X] = \sum_x xp(x)$
  - Continuous $E(X) = \int_x x\,P(X = x)$
  - Roll dice, flip coin, uniform distribution
- Population mean v.s. sample mean
  - Sample mean **itself** is random variable, estimator of population mean
  - Unbiased: E(sample mean) = population mean
- Why always larger sample is better?
  - Biased when fewer data?
  - Proof. Hint: E(X + Y) = E(X) + E(Y)

一个人去赌场，花5刀玩一个游戏，扔两次股子，如果和为6，他赢21刀，其他就什么也没有

问，这个游戏是偏向casino还是player的。（就是算个期望，然后期望是负的所以是偏向casino）

follow up: 如果现在有一个策略，这个人一直玩，直到第一次赢，然后走，问，他赢的概率是多少。

https://stats.stackexchange.com/questions/30365/why-is-expectation-the-same-as-the-arithmetic-mean

Variance

- Variance (σ^2) measure how random variable spreads

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

  - Calculate variance from rolling dice
- Standard deviation (σ)
- Sample variance (s^2) and standard deviation (s)

$$S^2 = \frac{\sum_{i=1}(X_i - \bar{X})^2}{n-1}$$

  - Why (n-1): Unbiased estimator of σ^2
  - Proof (HW), Hint Var(X+Y) = Var(X) + Var(Y) if X, Y independent

E(X+Y) = E(X) + E(Y); E(k*X) = k*E(X)

E(S^2) = Var(X)

Important discrete distribution

- Bernoulli
  - $P(X = x) = p^x(1-p)^{1-x}$
- Binomial
  - Sum of N Bernoulli variables
  - $$\mathrm{P}(X = x) \equiv \binom{N}{x} p^x(1-p)^{N-x}$$

E(X) = p, Var(X) = p (1-p)

What's the mean and variance?

E(X) = N*p, Var(X) = N*p(1-p)

Poisson: a given number of events occurring in a fixed interval of time

More details about other distributions

- Geometric distribution
  - Prob of seeing first success with k independent trials, each with success probability p.

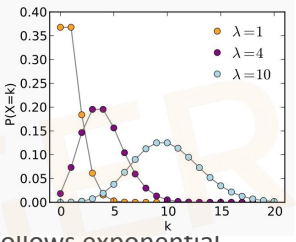
$$\Pr(X = k) = (1-p)^{k-1} p$$

- Negative binomial distribution
  - Prob of seeing r failures with k success trials , each with success probability p.

$$\Pr(X = k) = \binom{k+r-1}{k} p^k (1-p)^r$$

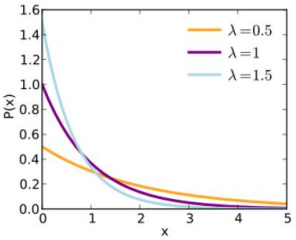
Important distribution

- Poisson, count data per time unit
  - $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$
  - Examples



- Exponential distribution
  - In a poisson distribution, inter-arrival time follows exponential distribution
  - PDF: $\lambda e^{-\lambda x}$ CDF: $1 - e^{-\lambda x}$
  - Expectation $E[X] = \frac{1}{\lambda}$
  - Variance $\mathrm{Var}[X] = \frac{1}{\lambda^2}$
  - "Memory-less"

    P(X > s+t | X > s) = P(X > t)



Proof of memoryless:http://stats.stackexchange.com/questions/2092/relationship-between-poisson-and-exponential-distribution

Really good read of binomial and poisson connection: https://probabilityandstats.wordpress.com/2011/08/18/poisson-as-a-limiting-case-of-binomial-distribution/

lamba is the mean arrival rate per unit time

1-e(-lambda * x)

Important continuous distribution

- Normal distribution

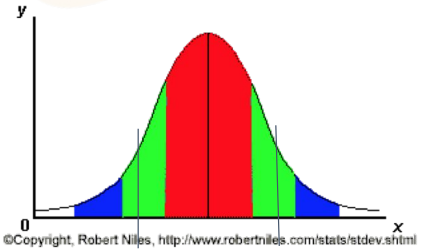
$$P(X = x) = (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}$$

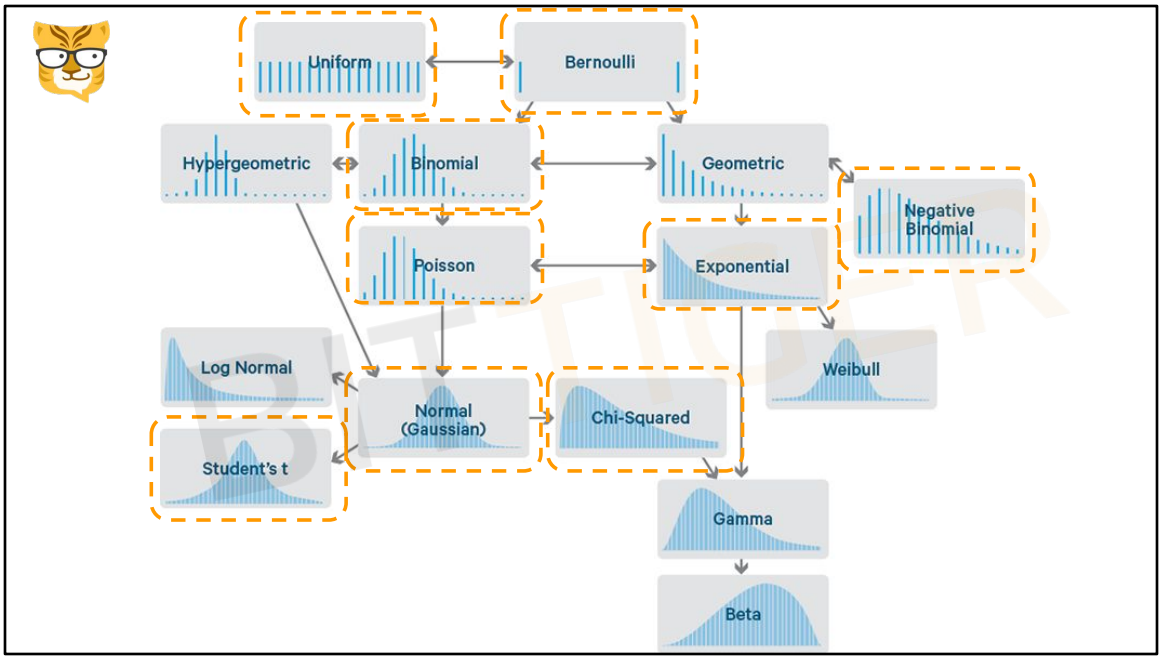
$$E[X] = \mu \text{ and } Var(X) = \sigma^2$$

- Standard normal distribution (Z) $\mu = 0$ and $\sigma = 1$

$$X = \mu + \sigma Z \sim \mathrm{N}(\mu, \sigma^2)$$

- Beauty of normal curve (6σ)
  - [μ - 3σ, μ + 3σ] covers 99.7%
  - [μ - 2σ, μ + 2σ] covers 95%
  - [μ - σ, μ + σ] covers 68%





http://blog.cloudera.com/blog/2015/12/common-probability-distributions-the-data-scientists-crib-sheet/
http://individual.utoronto.ca/zheli/poisson.pdf

Example interview questions

- In a country in which people only want boys every family continues to have children until they have a boy. If they have a girl, they have another child. If they have a boy, they stop. What is the proportion of boys to girls in the country?
  - What distribution is this?
  - What's expected number of kids (X) each family has?
  - What the ratio of boys to girls?

http://mathoverflow.net/questions/17960/google-question-in-a-country-in-which-people-only-want-boys

(Geometric) $P(X = k) = p(1-p)^{k-1} = (1/2)^k$, where k is the number of children a family has until they have first boy

$E(X) = Sum(k*P(X=k)) = 1*½ + 2* (½)^2 + ..$

$E(X) * ½ = 1 *(½)^2 + 2* (½)^3$

So E(X) = 2

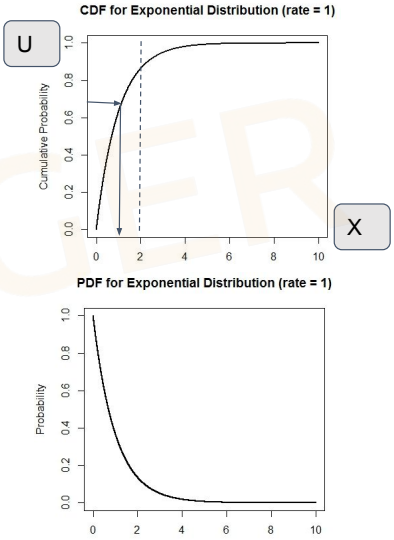
Then ratio is 50%

Another type of distribution Interview question

- How can we generate certain distribution?
  - Q1: Generate exponential distribution from uniform distribution
  - Q2: Generate normal distribution from uniform distribution

Inverse transform sampling

- To generate X with PDF: P(X)
  - Calculate CDF of X: F(X) and inverse $F^{-1}(.)$
  - Generate U with uniform distribution, say u
  - Calculate $F^{-1}(u)$
  - Repeat n times
- Intuition
- Let's solve Q1
- Can we solve Q2 as well?



Need to prove generated X = $F^{-1}(u)$ follows required PDF: P

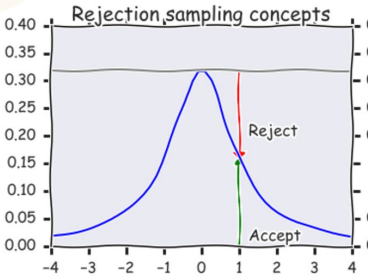
Meaning we needs to prove X follows required CDF: F, Let's see

CDF(X = x) = Prob(X <= x) = Prob($F^{-1}(u)$ <= x) = Prob(u <= F(x)) = F(x)

https://en.wikipedia.org/wiki/Inverse_transform_sampling
http://web.ics.purdue.edu/~hwan/IE680/Lectures/Chap08Slides.pdf

Acceptance rejection sampling

- Simple version: if good, accept; else, reject.
  - Ex1: Generate Y with uniform [¼, 1] given X ~ [0, 1]
  - Ex2: Generate random integer Y in [1, 7] given random integer X in [1, 5] 5*(X-1) + X
  - Ex3: Generate normal distribution given uniform distribution
    - Sample a point, x, calculate P(x)
    - Generate random number u from uniform distribution
    - If u < P(x), accept x, else reject, back to step 1
- Check out this video



https://en.wikipedia.org/wiki/Inverse_transform_sampling
(rand5() - 1) * 5 + rand5() gives us [1, 25], only accept [1, 21] then mod 7.

http://www.columbia.edu/~ks20/4703-Sigman/4703-07-Notes-ARM.pdf
http://www.tc.umn.edu/~elock/Rejection_Sampling.pdf
https://www.youtube.com/watch?v=tiWQwIISOuo

Picture from: http://people.duke.edu/~ccc14/sta-663-2016/15A_RandomNumbers.html

Percentile

- Definition (c%, N)

- Special percentiles
  - Median
  - 1st quantile
  - 3rd quantile
- Median, quantiles in normal distribution?
  - Look up **z score**, alpha: area outside [-z, z]

Why normal distribution is so important?

Law of large numbers

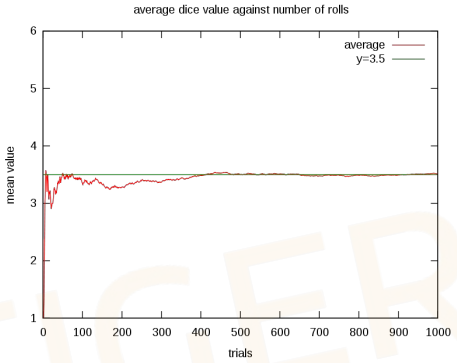
$X_1, X_2 \ldots X_n \ldots$ are independent, identically - distributed (IID) random variables, X has finite mean μ.

- Sample mean $\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$ converge to the true mean μ

  as n increases:

  <->

  $\bar{X}_n \xrightarrow[n\to\infty]{} \mu$



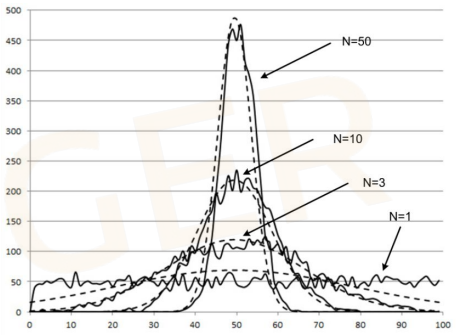
Extended material: http://www.math.uah.edu/stat/sample/LLN.html

Central limit theorem

- $X_1, X_2 \ldots X_n \ldots$ are independent, identically - distributed (IID) random variables, X has finite mean μ and variance σ^2

$$\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

(replacing σ by sample standard error, CLT still holds)



- Application
  - Binomial distribution
    - Flip coin example: $\bar{p} \sim N(p, \frac{p \cdot (1-p)}{n})$

The central limit theorem is a very important tool for thinking about sampling distributions – it tells us the shape (normal) of the sampling distribution, along with its center (mean) and spread (standard error)

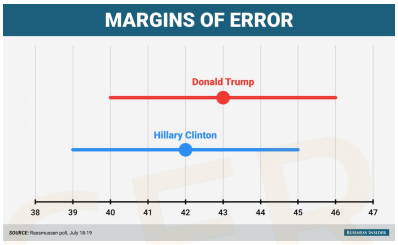
Confidence intervals

- Recall normal curve
  - What range covers 95% of possibility? 99%?
- Confidence interval (CI)
  - Derive 95% CI of observed $\bar{p}$ in Binomial distribution

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}}$$

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}} \qquad \hat{p} \pm \frac{1}{\sqrt{n}} \text{ approx.}$$

  - How many voters in the candidate poll?
  - Can Trump relax? How many are enough?



Fair coin p = ½ , what's 95% CI of **observed** p?

What if p is not known, use estimator to replace

When CLT doesn't hold?

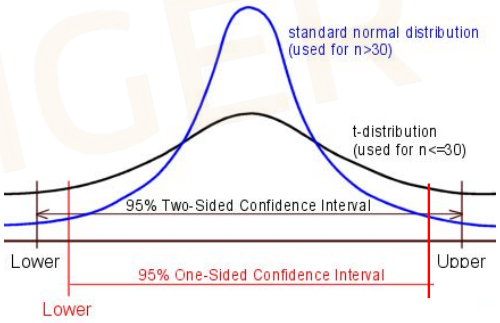
- Normal distribution -> T distribution when N < 30
  - T distribution has only one parameter: degree of freedom (df = N-1)
  - Approximate normal as df increases
  - CI under normal distribution

$$Mean_{estimate} \pm z_{1-\alpha/2} * StdErr_{estimate}$$

  - CI under t distribution

$$Mean_{estimate} \pm t_{n-1} * StdErr_{estimate}$$

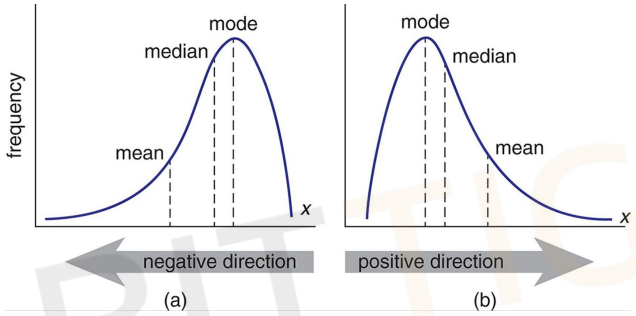
  - z or t?



t distribution was specially designed to provide more conservative test results when analyzing small samples (such as in the brewing industry).

What about CI of not normal distribution?

Skewed distribution



- What caused skewness
  - Bounded: lower bound or upper bound
- Can we still use normal distribution to estimate mean of skewed variable?

How to treat skewed variable

- Log transformation
- Use median, if so, what's CI of median?
  - Non-parametric method (v.s. parametric method)
    - **Bootstrap**
    - Jackknife

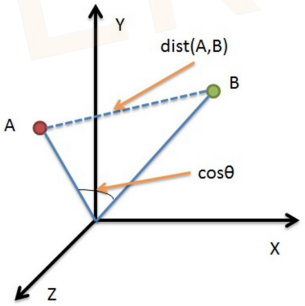
Covariance and correlation

- Definition
  - $cov(X, Y) = E[(X - E[X])(Y - E[Y])]$
  - $\varrho_{X,Y} = corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$
- What's sample covariance and correlation?
  - $cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} (\sum x_i y_i - n \bar{x} \bar{y})$
  - $cor(X, Y) = \frac{cov(X, Y)}{s_x s_y}$

If X= Y, E[(X - E(X)^2)]

Properties of correlation

- Properties
  - cor*(X, Y)* = cor*(Y, X)*
  - -1 <= cor*(X, Y)* <= 1
  - cor*(X, Y)* = 1 or -1: *X, Y* aligned on a line perfectly
  - *X, Y* are independent -> cor*(X, Y)* = 0
- Geometric interpretation
  - $$\varrho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i' y_i'}{\|x_i'\| \|y_i'\|} = cos(x', y')$$
  - cor*(X, Y)* = 0, cor*(Y, Z)* = 0 => cor*(X, Z)* = 0?



http://stats.stackexchange.com/questions/97051/building-the-connection-between-cosine-similarity-and-correlation-in-r