Foundations of Data Science

Lecture 1: Probability Theory and Statistics

MING GAO

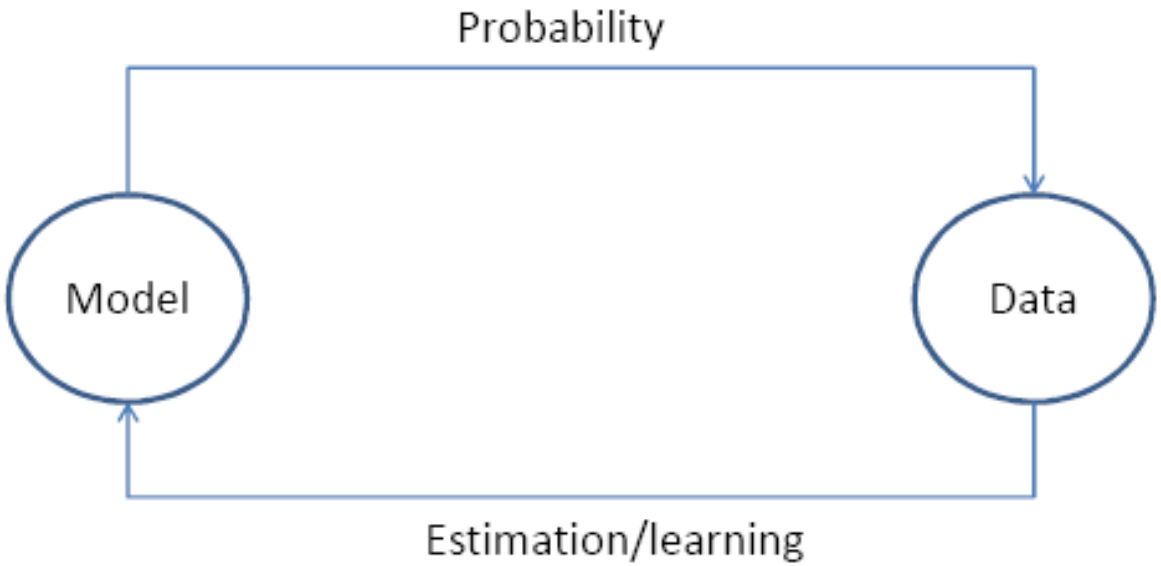
SE & DaSE @ ECNU
(for course related communications)
mgao@sei.ecnu.edu.cn

Sep. 18, 2016

Outline

1. Probability theory
   - Random variable
   - Joint probability distribution
   - Bayes rule
   - Moments
2. Statistics
   - Descriptive statistics
   - Estimation
   - hypothesis

Big picture of probability theory and statistics



Review

Definition

- Events and event space
- Random variables
- Joint probability distributions: marginalization, conditioning, chain rule, Bayes rule, law of total probability
- Moments

Sample space, events and event space

- Ω : sample space, result of an experiment
  - We toss a coin twice (Head = H, Tail = T)
  - $\Omega = \{HH, HT, TH, TT\}$
- Event is a subset of Ω
  - First toss is head, $E = \{HH, HT\}$
  - Two tosses are the same, $E = \{HH, TT\}$
- Event space, S, is a set of events
  - $S = \{s | s \subset \Omega\}$
  - Closed under finite union and complements: entails other binary operations, such as union, diff., intersection, etc.
  - Contains the empty event and Ω

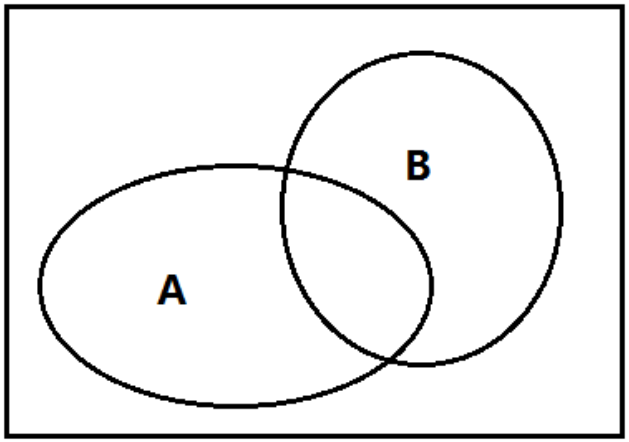
Probability measure

- Probability measure P is defined over (Ω, S), s.t.
  - $P(E) \geq 0$ for all $E \subset \Omega$.
  - $P(\Omega) = 1$
  - If E_1 and E_2 are disjoint, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$
- We can deduce other axioms from the above ones
  - $P(E_1 \cup E_2) =?$
  - $P(E_1 - E_2) =?$
  - $P(E_1 \cap E_2) =?$
  - $P(\overline{E}) =?$

Conditional probability

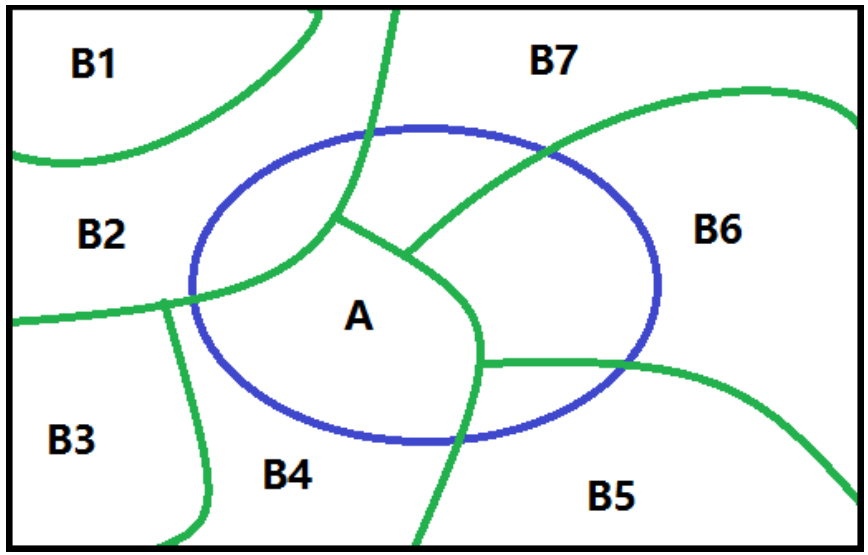
Definition

$P(A|B)$ is the fraction of worlds in which B is true that also have A true



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

The rule of total probability

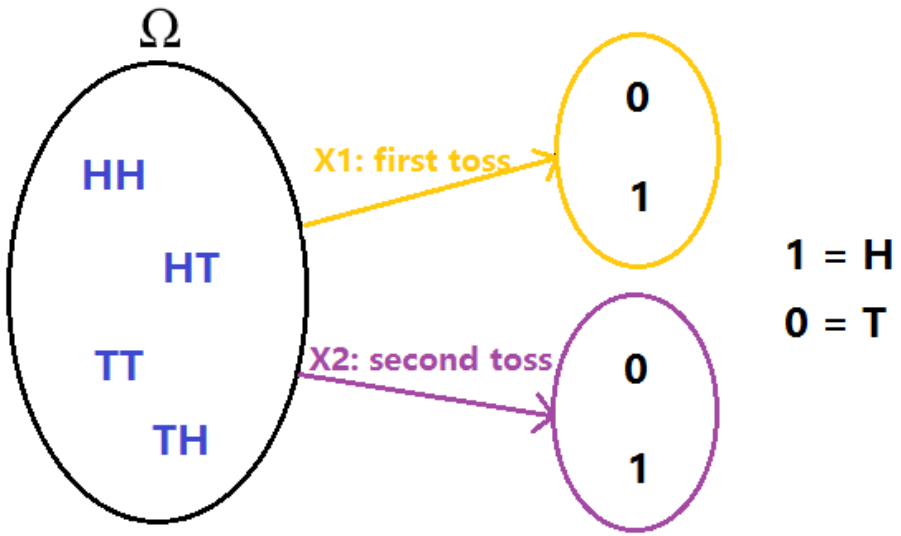


$$P(A) = \sum_i P(A|B_i) \times P(B_i) \tag{1}$$

From events to random variables

- Almost all the semester we will be dealing with random variables
- Modeling twice coin toss
  - $\Omega = \{HH, HT, TH, TT\}$
  - What are events?
  - Very cumbersome
- Concise way of specifying attributes of outcomes
- We need "functions" that maps from Ω to an attribute space

Random variable



- $P(X_1 = 1) = P(\{\text{the first toss is head}\}) = P(\{HH, HT\})$
- $P(X_2 = 0) = P(\{\text{the second toss is tail}\}) = P(\{HT, TT\})$
- $P(X_1 = 1, X_2 = 0) = P(\{HT\})$

Joint probability distribution

Given a set of random variables $X_1, X_2, \cdots$, where X_i denotes the outcome of the $i-$th coin toss and $X_i \in \{0, 1\}$

- We define a new r.v., Y, that denotes the number of tosses for the first head
  - $P(Y = n) = P(X_1 = 0, X_2 = 0, \cdots, X_{n-1} = 0, X_n = 1)$
  - If $P(X_i = 1) = p$, how to calculate the probability?
- A Bernoulli trial (or binomial trial) is a random experiment with exactly two possible outcomes, "success" (=1), "failure" (=0)
  - We fix the number of statistically independent Bernoulli trials to n
  - k denotes the number of successes, and $P(X_i = 1) = p$
  - How to calculate Binomial distribution $B(n, p)$?

Joint probability distribution Cont.

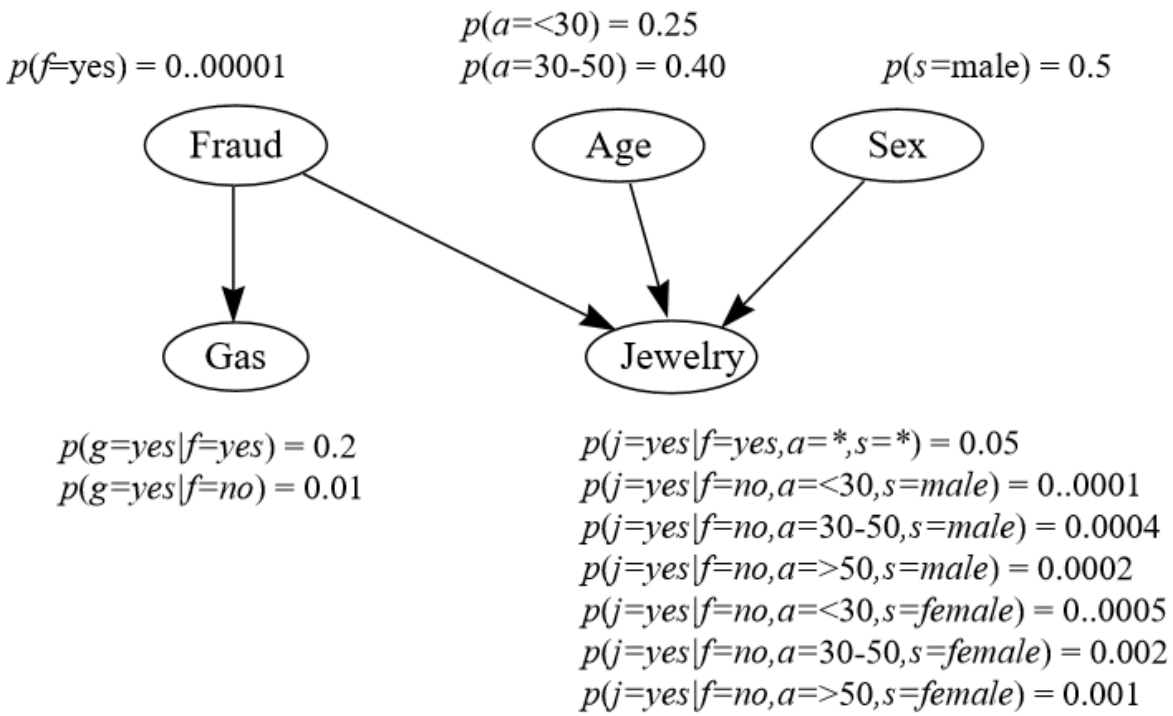
Given a set of random variables $X_1, X_2, \cdots, X_n$

- How to calculate probability $P(X_1 = x_1, X_2 = x_2)$
- How to calculate probability $P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$
  - Chain rule is always true
  -

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$$
$$= P(X_1 = x_1)\Pi_{i=2}^{n} P(X_i = x_i | X_{i-1} = x_{i-1}, \cdots, X_1 = x_1)$$

- Bayesian network
  $P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \Pi_i P(X_i = x_i | pa_i)$
- If X_i and X_j are independent,
  $P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \Pi_{i=1}^{n} P(X_i = x_i)$

Example of Bayesian network



$$P(a|f) = P(a), P(s|f,a) = P(s)$$
$$P(g|f,a,s) = P(g|f), P(j|f,a,s,g) = P(g|f,a,s)$$

Marginalization

- We know $P(X, Y)$, what is $P(X = x)$?
- We can use the low of total probability, why?

$$P(X = x) = \sum_y P(X = x|Y = y)P(Y = y)$$
$$= \sum_y P(X = x, Y = y)$$

- Another example

$$P(X = x) = \sum_{y,z} P(X = x|Y = y, Z = z)P(Y = y, Z = z)$$
$$= \sum_{y,z} P(X = x, Y = y, Z = z)$$

Bayes rule

- We know that $P(gender = 0) = 0.6$
- If we also know that the length of students hair is h_0, then how this affects our belief about her/his gender?

$$P(gener = 0|hair = h_0) = \frac{P(gender = 0)P(hair = h_0|gender = 0)}{P(hair = h_0)}$$

- Where does this come from?
- A simple naive Bayes model for Antispam
  - $S = 1$ and V denote that the email is a spam and it contains the word *viagra*
  -

$$P(S = 1|V) = \frac{P(V|S = 1)P(S = 1)}{P(V|S = 1)P(S = 1) + P(V|S = 0)P(S = 0)}$$

Naive Bayes classifier

- S denotes that the email is a spam
- We have collected a vocabulary, denoted as $X_1, X_2, \cdots, X_n$

$$P(S|X_1, X_2, \cdots, X_n) = \frac{P(X_1, X_2, \cdots, X_n|S)P(S)}{P(X_1, X_2, \cdots, X_n|S)P(S) + P(X_1, X_2, \cdots, X_n|\overline{S})P(\overline{S})}$$

- where $P(X_1, X_2, \cdots, X_n|S)P(S) = \Pi_{i=1}^{n} P(X_i|S)$ and $P(X_1, X_2, \cdots, X_n|\overline{S})P(\overline{S}) = \Pi_{i=1}^{n} P(X_i|\overline{S})$
- Why is this called Naive Bayes?
- To handle underflow, we calculate $\Pi_{i=1}^{n} P(X_i|S) = exp(\sum_{i=1}^{n} \log P(X_i|S))$.

Moments

X is a r.v.

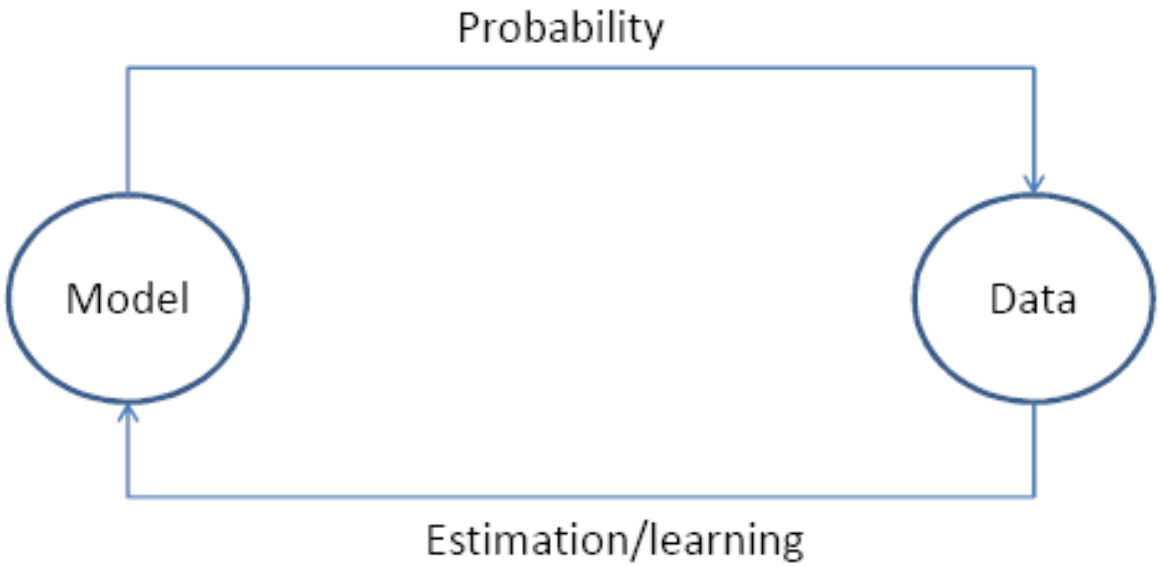
- Expectation: $\mu = E(X)$
  - Discrete r.v.: $E(X) = \sum_{v_i} v_i P(X = v_i)$
  - Continuous r.v.: $E(X) = \int_{-\infty}^{+\infty} x f(x) dx$
- Variance: $V(X) = E(X - \mu)^2$
  - Discrete r.v.: $V(X) = \sum_{v_i} (v_i - \mu)^2 P(X = v_i)$
  - Continuous r.v.: $E(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$
- For example $P(X = 1) = p$, and $P(X = 0) = 1 - p$
  - $E(X) = 1 \cdot p + 0 \cdot (1 - p) = p$
  - $V(X) = p(1 - p)^2 + (1 - p)(0 - p)^2 = p(1 - p)$

Properties of moments

X is a r.v.

- Expectation
  - $E(X+Y) = E(X) + E(Y)$
  - $E(aX) = aE(X)$
  - If X and Y are independent, $E(XY) = E(X)E(Y)$
- Variance
  - $V(X) = E(X^2) - [E(X)]^2$
  - $V(aX+b) = a^2 V(X)$
  - $V(X+Y) = V(X) + V(Y) + 2E(X - E(X))(Y - E(Y))$
  - If X and Y are independent, $V(X+Y) = V(X) + V(Y)$

Big picture of probability theory and statistics



Statistics Review

Overview

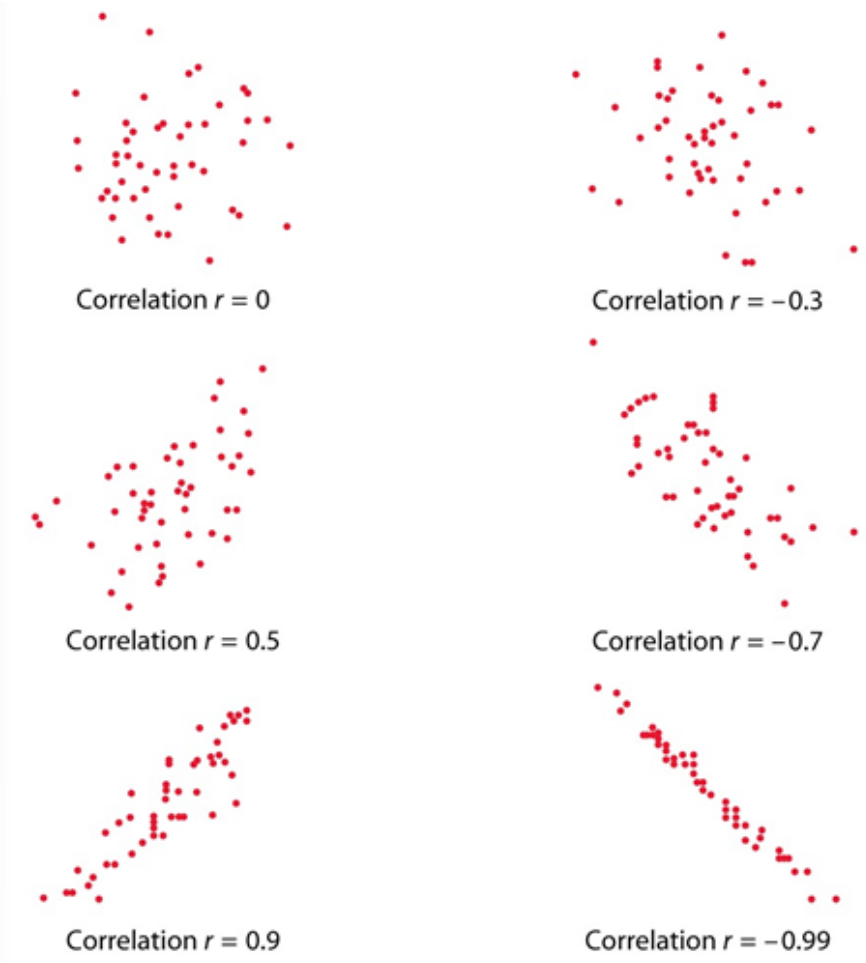
Statistics: Set of methods for collecting/analyzing data (the art and science of learning from data)

- Description: Graphical and numerical methods for summarizing the data
- Estimation: Learning parameters or distribution characteristics from data
- Inference: Methods for making predictions about a population (total set of subjects of interest), based on a sample (subset of the sample on which study collects data)

Numerical descriptions

- Let X denote a quantitative variable, with observations $x_1, x_2 \cdots, x_n$
- Describing the center
  - Mean: $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
  - Median: middle measurement of ordered sample
  - Mode: the value that appears most often in a set of sample
- Describing variability
  - Range: difference between largest and smallest observations
  - Variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$ (not n, due to technical reasons)
  - Standard deviation: $s = \sqrt{s^2}$
- Measures of position: p% quartile
  - p percent of observations below it, (100 - p)% above it
  - 50% quartile = median
  - Box plot: Minimum, 25% Q, Median, 50%, Maximum

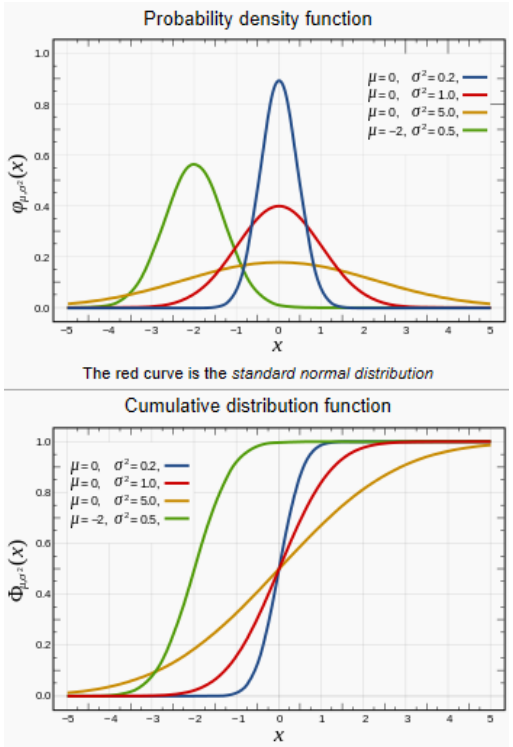
Correlation



Definition

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})\sum_{i=1}^{n}(y_i-\bar{y})}}$$

Normal distribution



Normal distribution

Most important probability distribution

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Symmetric, bell-shaped
- Characterized by mean (μ) and standard deviation (σ), representing center and spread
- An individual observation from an approximately normal distribution has probability
  - $P(\mu - \sigma \leq \mu + \sigma) \approx 0.68$
  - $P(\mu - 2\sigma \leq \mu + 2\sigma) \approx 0.95$
  - $P(\mu - 3\sigma \leq \mu + 3\sigma) \approx 0.997$

Central limit theorem

Theorem

For random sampling with "large" n, the sampling distribution of the sample mean $\overline{x}$ is approximately a normal distribution.

- $E(\overline{x}) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = E(X)$
- $Var(\overline{x}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i) = \frac{Var(X)}{n}$
- How "large" n needs to be depends on skew of population distribution, but usually $n \geq 30$ sufficient
- For example, you plan to randomly sample 100 students to estimate population proportion who have selected a course. $P(x = 1) = p$, $P(x = 0) = 1 - p$, $\overline{x} = \sum_{i=1}^{100} x_i$, then $\overline{x} \sim N(p, \frac{p(1-p)}{100})$.

Estimation

Goal

How can we use sample data to estimate values of population parameters?

- Point estimate: A single statistic value that is the best guess for the parameter value
  - Sample mean estimates population mean μ, $\widehat{\mu} = \overline{x}$
  - Sample std. dev. estimates population std. dev. σ, $\widehat{\sigma} = s$
  - Properties of good estimators: unbiased and efficient (smallest possible standard error)
- Interval estimate: an interval of numbers around the point estimate, that has a fixed "confidence level" of containing the parameter value. Called a confidence interval.
  - A confidence interval (CI) is an interval of numbers believed to contain the parameter value in form **point estimate $\pm$ margin of error**
  - For example, margin of error 2(standard error) for 95% confidence
  - $P(\mu - 1.96\widehat{\sigma} \leq \overline{x} \leq \mu + 1.96\widehat{\sigma}) \approx 0.95$, thus greater sample size gives narrower CI (we can further infer the sample size given margin of error)

Maximum likelihood estimation

Goal

Finding the parameter values that maximize the likelihood

- Suppose there is a sample $x_1, x_2, \cdots, x_n$ of n i.i.d. observations coming from a distribution with an unknown probability density function $f(\cdot)$.
- We first specifies the joint density function for all observations as $\mathcal{L}(\theta|x_1, x_2, \cdots, x_n) = f(x_1, x_2, \cdots, x_n|\theta) = \Pi_{i=1}^{n} f(x_i|\theta)$
- In practice it is often more convenient to work with the natural logarithm of the likelihood function, called the log-likelihood $\ln \mathcal{L} = \sum_{i=1}^{n} \ln f(x_i|\theta)$
- Maximum likelihood estimator (MLE) $\widehat{\theta} = \arg\max_{\theta \in \Theta} \ln \mathcal{L}(\theta|x_1, x_2, \cdots, x_n)$

MLE Cont.

Example

For a sample observing from a normal distribution

$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$

- $\ln \mathcal{L}(\mu,\sigma) = -\frac{n}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$
-

$$\frac{\partial}{\partial\mu}\ln \mathcal{L}(\mu,\sigma) = 0$$
$$\frac{\partial}{\partial\sigma}\ln \mathcal{L}(\mu,\sigma) = 0$$

- $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\widehat{\mu})^2$
- $E(\widehat{\mu}) = \mu$ and $E(\widehat{\sigma}^2) = \frac{n-1}{n}\sigma^2$

Expectation maximization algorithm

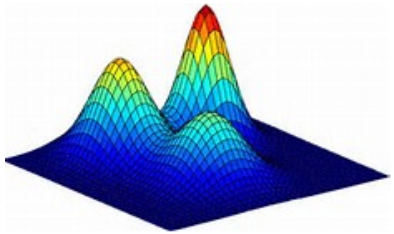
EM algorithm

The EM algorithm is used to find (locally) maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly

- Missing value
- Likelihood contain latent variables

- It consists of two steps: E-step (expectation) and M-step (maximization)
- It works in an iterative manner
- For example: Gaussian mixed model
  - Let $x = \{x_1, x_2, \cdots, x_n\}$ be a sample of n independent observations from a mixture of two multivariate normal distribution
  - Let $z = \{z_1, z_2, \cdots, z_n\}$ be the latent variables that determine the component from which the observation originates
  - $X_i|(Z_i = 1) \sim N(\mu_1, \Sigma_1)$ and $X_i|(Z_i = 2) \sim N(\mu_2, \Sigma_2)$

EM Cont.



GMM

- $P(Z_i = 1) = \tau_1,\ P(Z_i = 2) = \tau_2(1 - \tau_1)$, and $\theta = (\tau, \mu_1, \mu_2, \Sigma_1, \Sigma_2)$
- $f(x|\theta) = \sum_{i=1}^{2} \mathbb{I}_{z=j}\tau_i f(x|\mu_i, \Sigma_i)$
- $\ln \mathcal{L}(\theta; x, z) = \Pi_{i=1}^{n} \sum_{j=1}^{2} \mathbb{I}_{z_i=j}\tau_j f(x_i|\mu_j, \Sigma_j)$

E-step

$$T_{j,i}^{(k+1)} = P(Z_i = j|x_i; \theta^(k)) = \frac{P(Z_i = j, x_i|\theta^(k))}{P(x_i|\theta^(k))}$$

$$= \frac{\tau_j^{(k)} f(x_i|\mu_j^{(k)}, \Sigma_j^{(k)})}{\sum_{j=1}^{2} \tau_j^{(k)} f(x_i|\mu_j^{(k)}, \Sigma_j^{(k)})}$$

EM Cont.

E-step Cont.

$$Q(\theta|\theta^k) = E_{Z|X,\theta^k} \ln \mathcal{L}(\theta; x, z) = \sum_{i=1}^{n} E_{Z|X,\theta^k} \ln \mathcal{L}(\theta; x_i, z_i)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{2} P(Z_i = j|x_i; \theta^{(k)}) \ln \mathcal{L}(\theta_j; x_i, z_i) = \sum_{i=1}^{n}\sum_{j=1}^{2} T_{j,i}^k \ln \mathcal{L}(\theta_j; x_i, z_i)$$

M-step (for j = 1, 2)

$$\tau_j^{k+1} = \frac{\sum_{i=1}^{n} T_{j,i}^k}{\sum_{i=1}^{n}\sum_{j=1}^{2} T_{j,i}^k}$$

$$\mu_j^{k+1} = \frac{\sum_{i=1}^{n} T_{j,i}^k x_i}{\sum_{i=1}^{n} T_{j,i}^k}$$

$$\Sigma_j^{k+1} = \frac{\sum_{i=1}^{n} T_{j,i}^k (x_i - \mu_j^{k+1})(x_i - \mu_j^{k+1})^T}{\sum_{i=1}^{n} T_{j,i}^k}$$

Significance tests

Goal

We answer a question such as, "If the hypothesis were true, would it be unlikely to get data such as we obtained?"

- Spending money on other people has a more positive impact on happiness than spending money on oneself.
- Mental health tends to be better at higher levels of socioeconomic status (SES)

Definition

A significance test uses data to evaluate a hypothesis by comparing sample point estimates of parameters to values predicted by the hypothesis

- Null hypothesis (H_0): A statement that parameter(s) take specific value(s) (Usually: no effect)
- Alternative hypothesis (H_1): states that parameter value(s) falls in some alternative range of values (an effect)

Significance tests

Goal

We answer a question such as, "If the hypothesis were true, would it be unlikely to get data such as we obtained?"

- Spending money on other people has a more positive impact on happiness than spending money on oneself.
- Mental health tends to be better at higher levels of socioeconomic status (SES)

Definition

A significance test uses data to evaluate a hypothesis by comparing sample point estimates of parameters to values predicted by the hypothesis

- Null hypothesis (H_0): A statement that parameter(s) take specific value(s) (Usually: no effect)
- Alternative hypothesis (H_1): states that parameter value(s) falls in some alternative range of values (an effect)

Take-aways

- Probability review
  - r.v.
  - Probability operations
  - Bayes rule
- Statistics review
  - Descriptive statistics
  - Estimation
  - Hypothesis

Acknowledgement

Many slides are copied or adapted from:

- Prof. Anthony D. Joseph's slides for course cs109 (2006) at EECS@UCBerkeley (`http://cs109.github.io/2015/pages/videos.html`)