CE 395 Special Topics in Machine Learning

Assoc. Prof. Dr. Yuriy Mishchenko

Fall 2017

STATISTICAL LEARNING: PROBABILITY REVIEW

Why probability: statistical take on learning.

- How to learn from data?
  - **We have data** about a relationship, function, or algorithm
    - Example:
      - Data about customer's shopping activity,
      - Data containing images of cats and dogs,
      - Recordings of people talking about things,
      - Numerical solution of a complicated system of equations describing a nano-scale device
  - **Learn** that relationship, function, or algorithm **from data**
    - Example of learned things:
      - Predicting of what customer will shop for next,
      - Identifying cats or dogs in images,
      - Transcribing human speech into text,
      - Predicting device's behavior from its design parameters
  - **Answer to this question?** %-x

Statistical take on learning

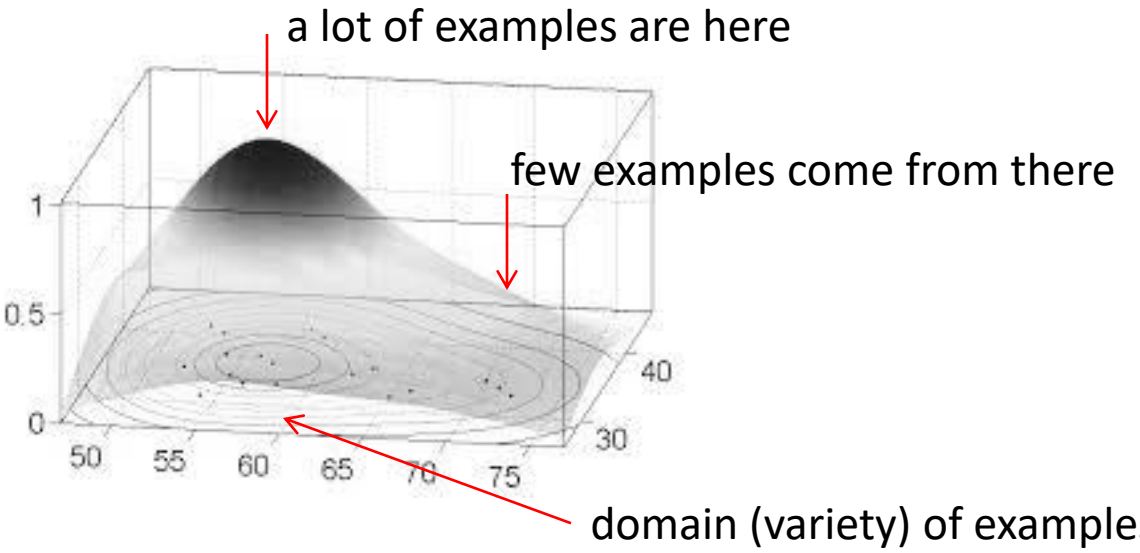
A shift in the point of view:

- **What we learn in programming courses:** design a solution (for relationship, function, algorithm) manually by analyzing and thinking about problem first and then devising steps towards the solution
- **Where does this fail:** algorithms that are complex beyond comprehension – for example, transcribe human speech
- **What "learning" suggests:** learn solution to problem directly from data

Statistical take on learning

- Learn from data… **What is data?**
  - P(X) – collection of examples such as cat or dog images
  - P(X,Y) – relationship instances such as “if X then Y”
- What is P?
  - For a collection of examples, P describes the variety of examples possible to see as well as how many of each example we may have
  - We will call P a **distribution** for now

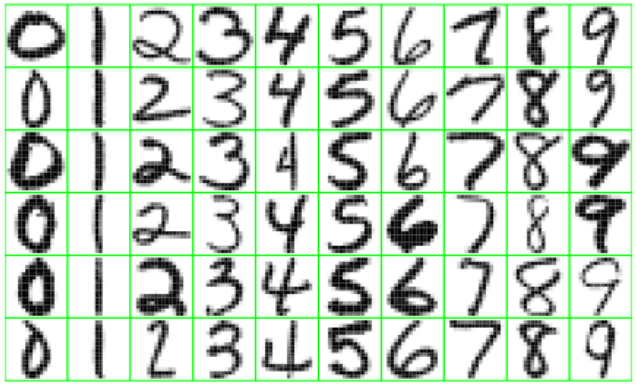
P – a distribution:

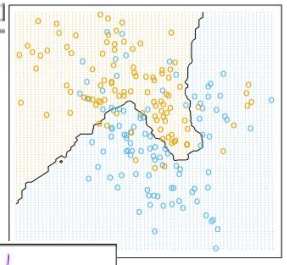


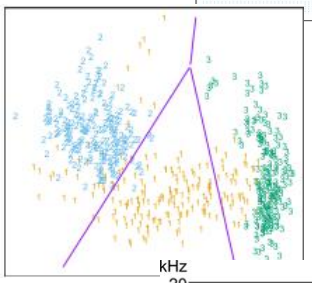
Statistical take on learning

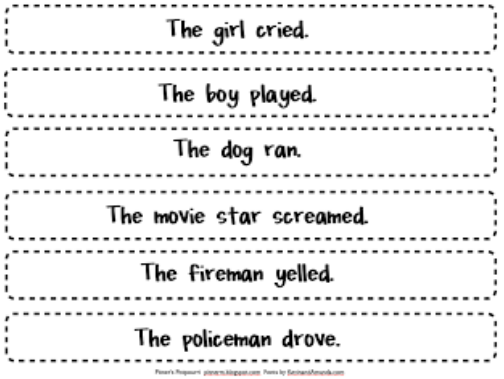
Think about data and P(X) for a moment, and how they are related.

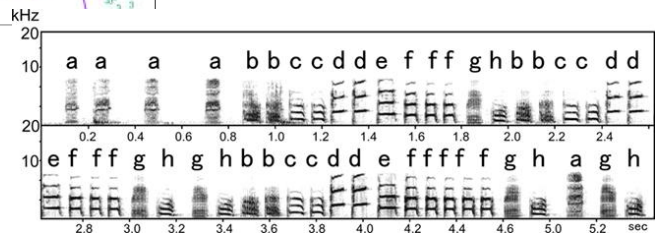
	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

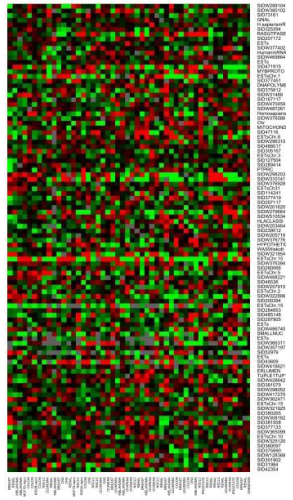












ALL OF THESE EXAMPLES ARE DATA

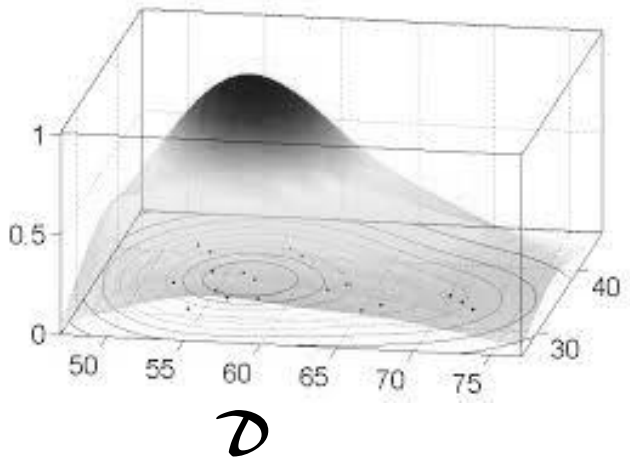
Statistical take on learning

Can the notion of "distribution" P(X) describe arbitrary "data" there can be?

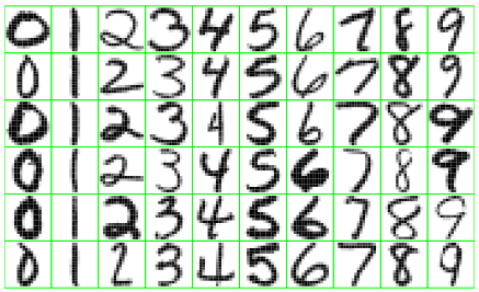
- The answer is Yes

Arbitrary data can be abstracted as a domain set of examples $X \in \mathcal{D}$ and the distribution over the examples P(X)

P(X)



$\mathcal{D}$



Statistical take on learning

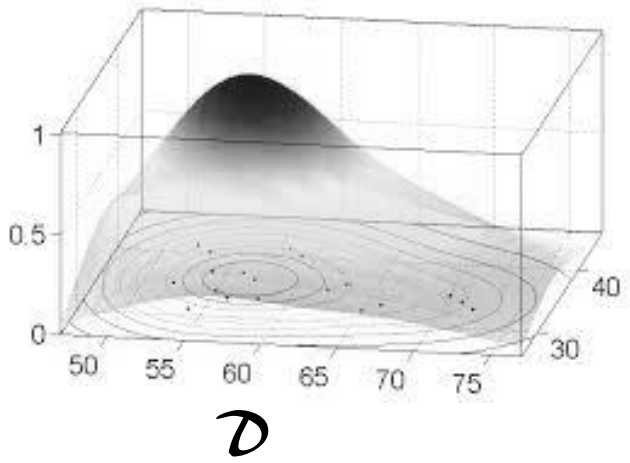
Does knowing P(X) describes everything there is to know about "data"?

- The answer is also Yes

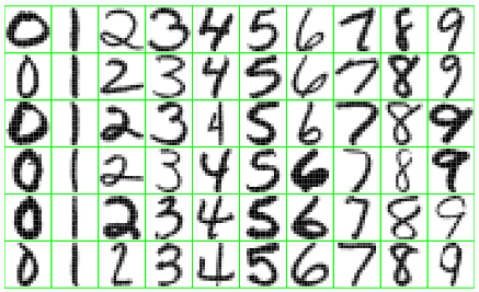
Generating datasets from P(X):

- Select X randomly from P(X) = Data{X_i}

P(X)

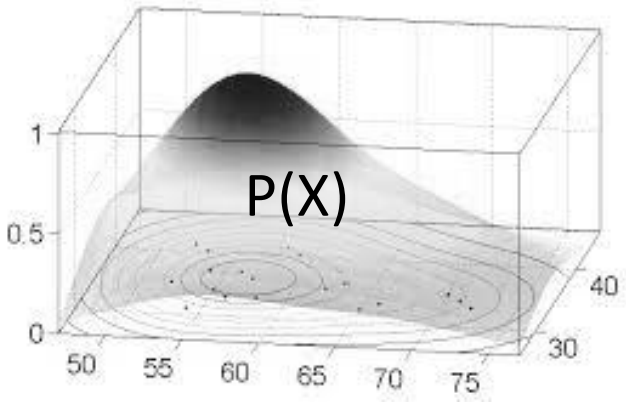


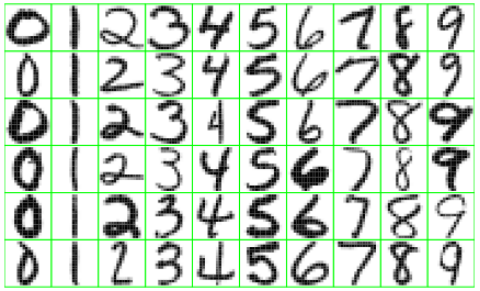
$\mathcal{D}$



Statistical take on learning

P(X)=DATA





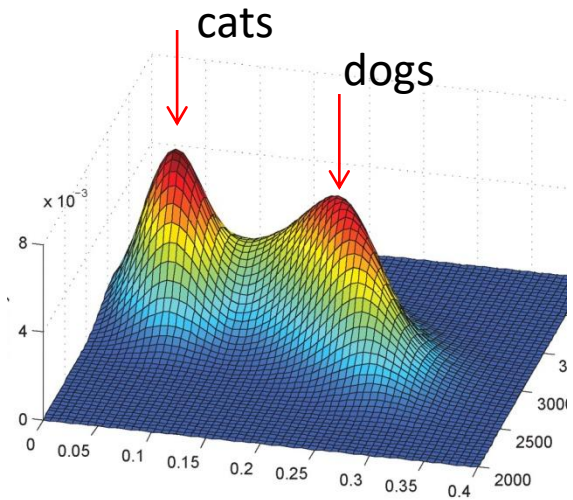
Statistical take on learning

Statistical take on learning:

- Learn the distribution P(X) from a series of its examples $\{X_i\}$

$\{X_i\}$

P(X)

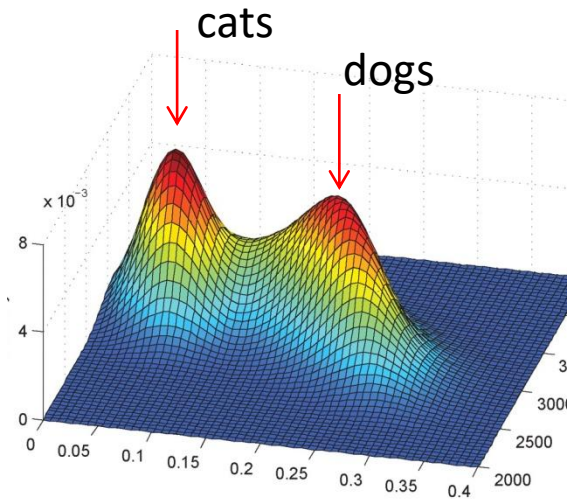


Statistical take on learning

For example, let's think about learning P(X) itself given a number of examples from it, X. This is **unsupervised learning** or **clustering**: Concentrations or clusters in such P(X) typically imply entities (eg cats vs. dogs if X were images)

$\{X_i\}$

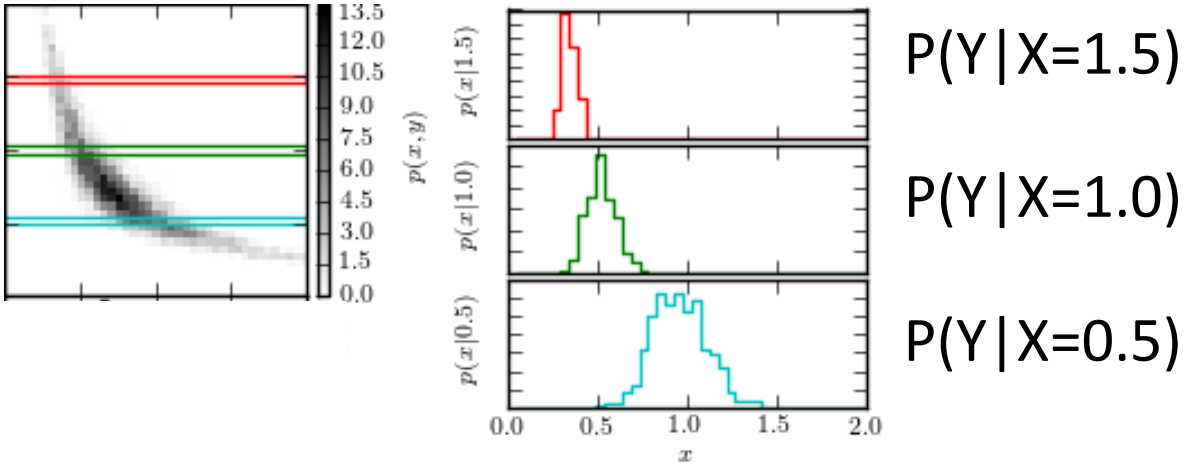
P(X)



Statistical take on learning

Learning P(X,Y) can be used also for **relationships, functions** and for **prediction** ("if X then also means Y")

P(X,Y)



P(Y|X=1.5)

P(Y|X=1.0)

P(Y|X=0.5)

Statistical take on learning

How does machine learning fit in?

- ML is learning P(X) through **models** $\pi(X)$
  - Models may be **parametric** such as $\pi(X;\theta)$
  - Or models may be **nonparametric** such as K-nearest neighbors $\pi(X;\{X_i\})$ (depends on data directly without parameters)

Statistical take on learning

Statistical learning is not the only possibility for learning in AI, other kinds of learning exist:

- Logic and expert systems
- Reinforced learning
- Etc.

Statistical take on learning

The key notions from probability that we will needed for statistical learning:

- Random variables
- Probability distributions
- Expectation values
- Variance, covariance, and correlations
- Averages and the Law of Large Numbers (LLN)

Our discussion of these will be informal

Random variables

- **Random variable** is a variable whose value is outcome (usually numerical) of a random process or phenomenon
- We typically write (capital) X to represent random variables or r.v.
- Note that X is not a value by itself; one needs to "inspect" X to get a value, and every time "inspection" happens X is different!
- We write x to represent the actual values of "looking" at X – (those are random)

Random variables

Examples:

- Dice throw X is a random variable that takes on randomly the values 1, 2, 3, 4, 5, 6 every time we look at it

Random variables

Examples:

- Coin throw X, is a random variable that takes on randomly the values “heads” and “tails” every time we look at it (also can be 0 or 1)

Random variables

Now you :

- Double-dice throw …

Random variables

Random variables don't have to be simple (or numeric for that matter) and can represent really complex things:

- Time it takes to arrive from F.Altay to IEU, T
- Image randomly drawn from a database of hand-written letters, I
- Purchases made by customers, C
- Next president of the United States, P

Random variables

- While X itself is not a value, it may be part of a relationships that is understood to be true for any realization of X
- For example:
  - $1 \leq X_{dice} \leq 6$ is always true no matter which X_{dice} we had
  - $X < X+1$ is always true no matter what X is
  - $Y = X^2$ here two random variables are “linked”, that is, they change precisely together albeit at random
  - $X+Y=0$ (same)

Random variables

Example:

- Let's say you arrive to F.Altay to take a bus to IEU always at 8:30 am
- Let us define two r.v.: the time it takes to reach IEU from F.Altay – X, and whether you are late for your class at 9:00 a.m. – Y
- X and Y are random as they change every day. Yet, without knowing any actual X and Y we can claim **X>30 min $\Rightarrow$ Y="Late"**
- This is a random relationship: X and Y change randomly every day; however, the above is **always true** no matter which day you look at it – if it took 15 minutes to reach IEU today, you were on time; if it took 35 minutes to reach IEU yesterday, you were late.
- **X>30 min $\Rightarrow$ Y="Late"** means that X and Y are related and although each takes on random values, those values never violate the relationship!

Random variables

Sample space is the set of all possible outcomes of random variable inspections (can be **discrete**, **continuous**, or **mixed**)

- Sample space of r.v. for dice is {1,2,3,4,5,6}
- Sample space of r.v. for coin toss is {0,1}
- Sample space of r.v. for 2 dices is 1..12
- Sample space of r.v. representing the time it takes bus to arrive from F.Altay to IEU is $(0,\infty)$ (continuous)

Probability distribution

- P(X=x) is called **probability distribution**, **probability measure**, **probability density function**, or **probability mass function**; P{X=x}, Pr{X=x}, P{x} and Pr(x) are other frequent notations
- P(x) represents the relative number of outcomes in which r.v. X takes value x
- Note that P(x) depends on the **value** of r.v. X
- We write P(X) to indicate the probability distribution of the entire r.v. X, as in “for all values of X”

Random variables

Example:

- For dice throw X, (P{X=1}=) P(1)= P(2)= P(3)= P(4)= P(5)= P(6)= 1/6

Random variables

Example:

- For coin toss X, P(“heads”)=P(“tails”)=1/2

Random variables

Now you :

- For 2 dices throw, P(x)=

x	1	2	3	4	5	6	7	8	9	10	11	12
P(x)												

Dice1	Dice2	Total
1	1	
2	2	
3	3	
4	4	
5	5	
6	6	

Expectation values

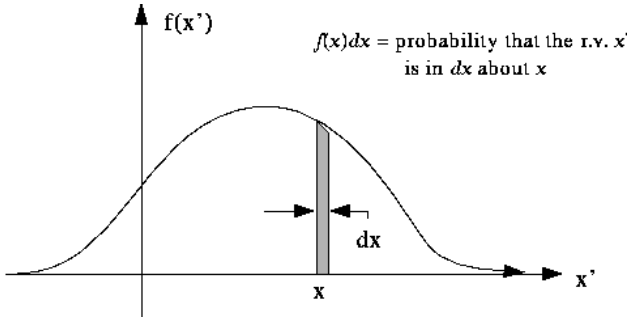
Key property of probability: probability of **any** outcome is 100% always:

$$\Pr\{x\,is\,anything\} = \sum_{all\ \ x} P(x) = 1$$

Probability distribution

- If X is continuous, any individual value of X has the probability of 0 to be observed;
  - For example, consider X that uniformly takes real values 0<X<1. What is the probability of seeing X at value x = 0.172827336261… ?
- Instead, we talk about the probability of X being inside a small interval $[x, x+dx]$, which is $\Pr\{[x, x+dx]\}=p(x)dx$
- p(x) therefore is called **probability density function**

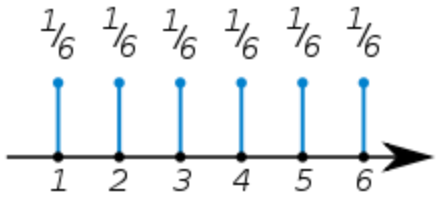
probability density function



Probability distribution

- If X has a discrete set of possible values, P(X= x) is finite for any single x and is often called **probability mass function**, as in P(x) is a finite probability "mass" (eg 1/6) attached to point-like possible outcomes of X

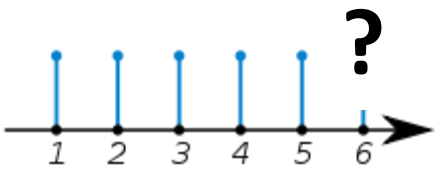
probability mass function



Probability distribution

- P(X) may feature a mix of mass and density:
- Consider a random number generator where you first roll a 6-value dice and print the dice value if it was smaller or equal than 5. If dice=6, then you run a continuous random number generator on your computer that produces a number between [0,1], and print that value + 6 instead. What is the probability distribution of that r.v. X?

Represent graphically this probability distribution:



Expectation values

- A particularly important role in probability is played by **expected or expectation values**
- Expected value of r.v. X is defined as (definition)

$$E[X] = \sum_x x \cdot P(x)$$

$$E[X] = \int x \cdot p(x) dx$$

Expectation values

The key property of expectation values:

- $E[X+Y]=E[X]+E[Y]$
- $E[\sum X_i]=\sum E[X_i]$ **(always true)**

Expectation values

Examples:

$$E[X] = \sum_{x=1}^{6} x_i \frac{1}{6} = 3.5$$

x	1	2	3	4	5	6
P(x)	1/6	1/6	1/6	1/6	1/6	1/6

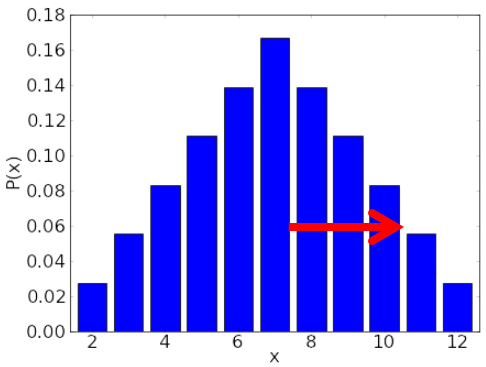
$$E[X_{1+2}] = E[X_1 + X_2] =$$

x	1	2	3	4	5	6	7	8	9	10	11	12
P(x)	1/36	2/36	3/36	4/36	5/36	6/36	6/36	5/36	4/36	3/36	2/36	1/36

Descriptive statistics

- **Mean of r.v. X** is defined as **m=E[X]**=$\sum$xP(x); mean represents the "central value" of X
- **Variance of r.v. X** is defined **S=var(X)**=E[(X-m)^2]= $\sum$(x-m)^{2}P(x) and represents the "spread" of X around its central value
- $\sigma=\sqrt{S}$ is called **standard deviation** and likewise represents the spread of X around central value

Two dice throw outcome:



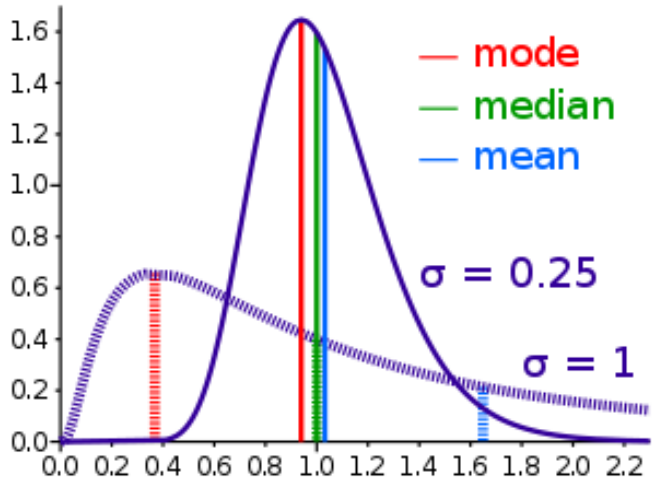
STD - spread

Descriptive statistics

One can show rather easily $S=E[X^2]-E[X]^2$

Descriptive statistics

- **Mode of r.v. X or probability distribution P(X)** is the point where P(x) reaches maximum
- **Median of X or P(X)** is defined as such a value *a* that Pr{X<*a*}=50%



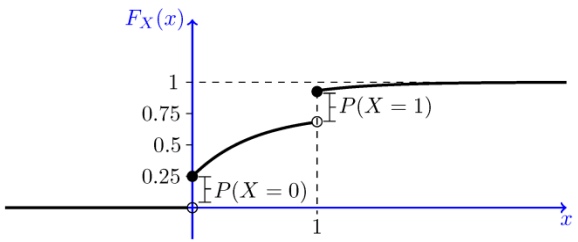
Cumulative probability distribution

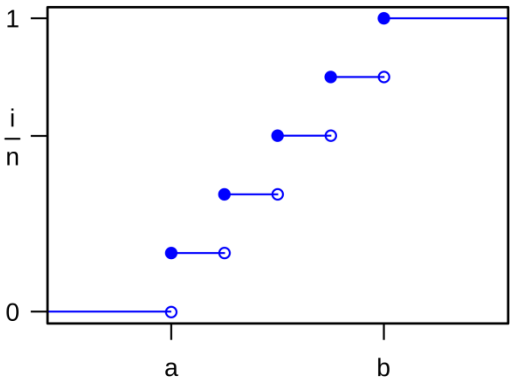
- If **probability distribution function (PDF)** P(X=x) indicates the probability of the event that having looked at X we saw x, the **cumulative distribution function (CDF)** indicates the probability that X will take a value smaller than x:

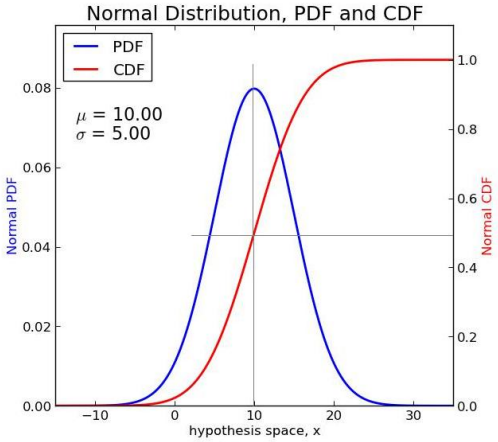
$$F(x) = \Pr\{X \leq x\}$$

Cumulative probability distribution

$$F(x) = \sum_{x' \le x} P(x') = \int_{-\infty}^{x} dx'\, p(x')$$

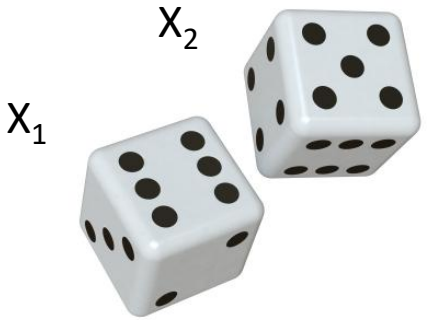






Joint probability distribution

- P(X,Y) is called a **joint probability distribution** and means the probability of X=x and Y=y occurring at the same time.
- Joint probability distribution can be visualized as a table. For two dices $P(X_1,X_2)$ is $\rightarrow$



X1/X2	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

Joint probability distribution

Note:

$$\sum_{all\ x}\sum_{all\ y} P(x,y) = 1$$

CHECK THIS →

X1/X2	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

Marginal distributions

- If we want to calculate E[X] using P(X,Y), what shall we do?
  $\mathrm{E}[X]=\sum_{x,y} x\mathrm{P}(x,y)= \sum_x x\sum_y \mathrm{P}(x,y)=\sum_x x\cdot$**[WHAT?]**
- $\sum_y \mathrm{P}(x,y)$ is called **marginal distribution** and is exactly Pr{X=x}: $\mathrm{P}(x)=\sum_y \mathrm{P}(x,y)$, $\mathrm{P}(y)=\sum_x \mathrm{P}(x,y)$

P(D,E)

Degree	Work Experience			
	5 years or less	10 years	15 years	Total years
None	.05	.09	.16	.30
Junior College	.03	.10	.13	.27
Big Ten	.10	.08	.04	.23
Ivy League	.16	.03	.01	.20
	.35	.30	.35	1.00

P(D)

P(E)

Conditional probability

- **Conditional probability** is the probability of observing Y *if X was already observed*, denoted P(Y|X)

Degree	Work Experience			
	5 years or less	10 years	15 years	Total years
None	.05	.09	.16	.30
Junior College	.03	.10	.13	.27
Big Ten	.10	.08	.04	.23
Ivy League	.16	.03	.01	.20
	.35	.30	.35	1.00

- **P(Y|X)=P(X,Y)/P(X)** follows from the fact that must have $\sum_y P(y|x)=1$ –it is a probability after all!

Conditional probability

- The latter statement is known as the **chain or the product rule**:

$$\sum_y P(y \mid x) = 1 \Rightarrow P(y \mid x) = \frac{P(x,y)}{\sum_y P(x,y)} = \frac{P(x,y)}{P(x)}$$

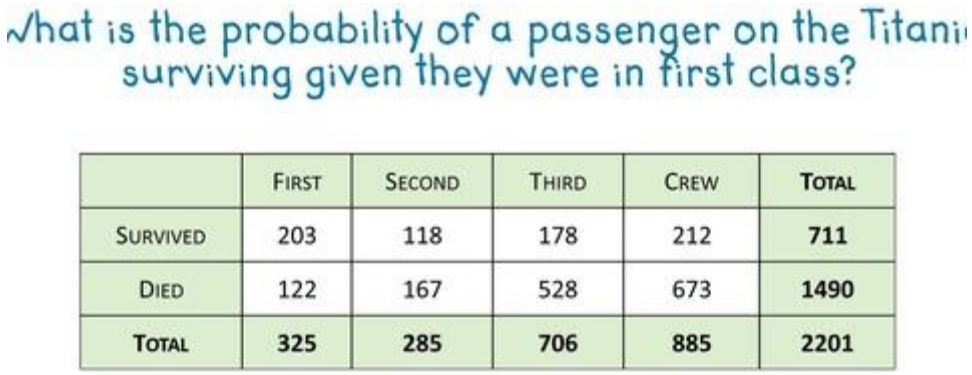
$$OR \; P(x,y) = P(y \mid x)P(x)$$

CHECK THE CHAIN RULE HERE

Degree	Work Experience			
	5 years or less	10 years	15 years	Total years
None	.05	.09	.16	.30
Junior College	.03	.10	.13	.27
Big Ten	.10	.08	.04	.23
Ivy League	.16	.03	.01	.20
	.35	.30	.35	1.00

Conditional probability

Use conditional probability to answer this question:



	First	Second	Third	Crew	Total
Survived	203	118	178	212	**711**
Died	122	167	528	673	**1490**
Total	**325**	**285**	**706**	**885**	**2201**

$$P(Y \mid X) = \frac{P(X, Y)}{P(X)}$$

Rules of probability

- Sum Rules

$$1 = \sum_{x} P(x)$$

$$P(x) = \sum_{y} P(x, y)$$

$$P(x_1) = \sum_{x_2} \sum_{x_3} \ldots \sum_{x_n} P(x_1, x_2, \ldots, x_n)$$

- Product/Chain Rule

$$P(x, y) = P(y \mid x) P(x)$$

$$P(x_1, x_2, \ldots, x_n) = P(x_1) P(x_2 \mid x_1) \ldots P(x_n \mid x_1, \ldots, x_{n-1})$$

Bayes Theorem

The following uncharacteristically important for its simplicity formula is known as the **Bayes Theorem**

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \Leftarrow$$

$$P(X, Y) = P(Y \mid X)P(X) = P(X \mid Y)P(Y)$$

Bayes Theorem

Bayes theorem allows one to "reverse" causality in probability:

- Lets' say that X makes Y, X$\rightarrow$Y
- Say now we observed Y, what is the probability of X, P(X|Y) ?

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

Bayes Theorem

Example:

$$P(sun \mid hot) = \frac{P(hot \mid sun)P(sun)}{P(hot)}$$

Conditional Distributions

$P(W|T)$

$P(W|T = hot)$

W	P
sun	0.8
rain	0.2

$P(W|T = cold)$

W	P
sun	0.4
rain	0.6

Joint Distribution

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Bayes Theorem

Example:

$= 8$

$$P(X_1 = 5 \mid X_1 + X_2 = 8) = ?$$

$$P(X_1 = 5 \mid X_1 + X_2 = 8) = \frac{P(X_1 + X_2 = 8 \mid X_1 = 5) P(X_1 = 5)}{P(X_1 + X_2 = 8)} = \frac{1/36 \cdot 1/6}{5/36} = \frac{1}{5}$$

Independence

- Two r.v. X and Y are called **independent** if $P(X,Y)=P(X)P(Y)$ or, equivalently, $P(X|Y)=P(X)$ and $P(Y|X)=P(Y)$

- Note the equivalence $P(X|Y)=P(X) \Leftrightarrow P(Y|X)=P(Y) \Leftrightarrow P(X,Y)=P(X)P(Y)$ – is easy to show by the **chain rule** ! (**Try to do that yourself**)

Independence

- Two r.v. X and Y are conditionally independent given Z if they are independent after Z is fixed, in the conditional distribution P(X,Y|Z)
- This implies P(X,Y|Z)=P(X|Z)P(Y|Z).
- Corollary: **P(X,Y,Z)=P(X|Z)P(Y|Z)P(Z)**

Example: [whiteboard]

Some important technicalities

The expectation value E is a linear operation, that means you can switch orders of E with any other linear operation, so

- $E[X+Y]=E[X]+E[Y]$
- $E\Sigma Xi=\Sigma Exi$
- $E\int f(X)=\int Ef(X)$
- $E[Filter[X]]=Filter[E[X]]$
- $E[h*X]=h*E[X]$

Some important technicalities

The same cannot be said about E and the product ($\cdot$), E and $\cdot$ cannot generally be interchanged $E[XY] \neq E[X]E[Y]$ – try any example for yourself.

However, **if X and Y are independent, THEN $E[XY]=E[X]E[Y]$** – this is a very important property of independent r.v. that you surely **must remember**!

Some important technicalities

Another important property of independent variables, related to variance, is

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y)$$

$$\mathrm{var}(\sum X_i) = \sum \mathrm{var}(X_i)$$

Try and show this from P(X,Y)=P(X)P(Y)

Important: this property is only true if r.v. are independent, in general var(∑X)≠∑var(X) !

Some important technicalities

- A **moment of order n** of r.v. X, $m_n(X)$, is $E[X^n]$. Thus, $m_1(X)=E[X]$ is the expectation value and $m_2(X)=E[X^2]=E[X]^2+\text{var}(X)$.
- The latter identity is very important for you to know, here, I write it once again:

$$E[X^2] = E[X]^2 + \text{var}(X)$$

Some important technicalities

- For two variables, a quantity related to E[XY] is called covariance.
- More precisely, **covariance is defined as**

$$\mathrm{cov}(X,Y) = E[(X - E[X])(Y - E[Y])$$

- Following identity is important and also relatively easy

$$\mathrm{cov}(X,Y) = E[XY] - E[X]E[Y]$$

Some important technicalities

Covariance quantifies the degree of "relatedness" of two r.v.

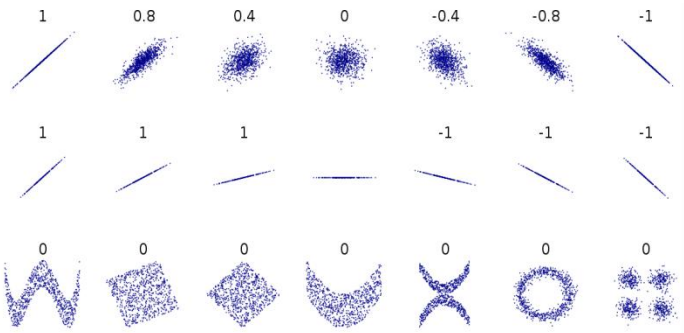
- If X and Y are independent, then $E[XY]=E[X]E[Y]$ and $cov(X,Y)=0$.
- If X and Y are perfectly related, $X=aY$, then $cov(X,Y)=a\ var(X)$ and is maximal
- If X and Y are perfectly anti-related, $X=-aY$, then $cov(X,Y)=-a\ var(X)$ and is minimal

Some important technicalities

- The ratio of covariance to the geometric mean of r.v. variances is called **correlation**

$$corr(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}} \in [-1,1]$$

- **Correlation** has the same meaning as the covariance, just it is rescaled to be in the interval [-1,1]



Some important technicalities

Knowing covariance, we can now readily write down the general relationship of the variance of sum of r.v.

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\,\mathrm{cov}(X,Y)$$

$$\mathrm{var}\left(\sum X_i\right) = \sum_i \mathrm{var}(X_i) + 2\sum_{i<j} \mathrm{cov}(X_i, X_j)$$

Now, you know why the variances of independent r.v. add together – because the covariance of such r.v. is zero !

Some important technicalities

For several r.v. X_i, the covariance matrix Σ is defined as

$$\Sigma_{ij} = \mathrm{cov}(X_i, X_j)$$

Some important technicalities

Example calculating the covariance matrix:

$$\Sigma_{ij} = E[(X_i - E[X_i])(X_j - E[X_j])] = E[X_i X_j] - E[X_i]E[X_j]$$

X_1 / X_2	1	2	3
1	0.1	0.1	0.1
2	0.2	0.1	0.1
3	0.1	0.2	0

Σ	x_1	x_2
x_1		
x_2		

$E[X_1] = 2$

$E[X_2] = 1.8 \quad E[X_1^2] = 4.6,\ E[X_2^2] = 3.8,\ E[X_1 X_2] = 3.5$

Some important technicalities

- In many cases we will be interested in arithmetic averages (or simply averages) of independent random variables

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Note that S_n is actually a random variable itself – its value is tied to the realizations of random variables X_i !

Some important technicalities

- It is fairly easy to conclude that the expected value of S_n is the average of the expected values of X_i

$$E[S_n] = \frac{1}{n}\sum_{i=1}^{n} E[X_i]$$

- However, a more difficult question is how far from that average the value of S_n can deviate (since it is a random variable) ?!

Some important technicalities

- An extremely important result in probability is known as the **Law of Large Numbers** (LLN, also known as Chebyshev inequality, see Advanced).
- LLN states that as n grows, nearly 100% of the possible values of S_n will be concentrated at $E[S_n]$, without any significant deviations!

$$S_n \to \frac{1}{n}\sum_{i=1}^{n} E[X_i],\ as\, n \to \infty$$

Some important technicalities

- Formally, it is said that S_n converges *in probability* to ave(E[Xi]). That is, as n grows, the probability to find S_n deviate from ave(E[Xi]) by any significant amount δ goes down as 1/n.

$$\Pr\{|S_n - \frac{1}{n}\sum_{i=1}^{n} E[X_i]| \geq \delta\} \leq \frac{\text{ave}(\text{var}(X_i))}{n\delta^2}$$

Some important technicalities

- The importance of LLN is that it allows to reasonably replace averages of large numbers of random variables with a single number – the mean of their expectation values:

$$S_n \overset{r.v.}{\to} \frac{1}{n}\sum_{i=1}^{n} E[X_i],\ as\ n \to \infty$$

- This fact is extremely general, and only depends on that the summed X_i are independent and the variances var(X_i) are bounded

Some important distributions

- **Bernoulli** – represents outcome of a biased coin toss $X \in \{0,1\}$, $P(X=1)=p$

$$P(x) = p^x (1-p)^{1-x}$$

- **Poisson** – represents the count of events (for example phone calls to a call center) arriving over time with constant rate r

$$P(x) = \frac{r^x}{x!} e^{-r}$$

- **Binomial** – represents the number of Bernoulli outcomes after N coin tosses

$$P(x) = \frac{N!}{x!(N-x)!} p^x (1-p)^{N-x}$$

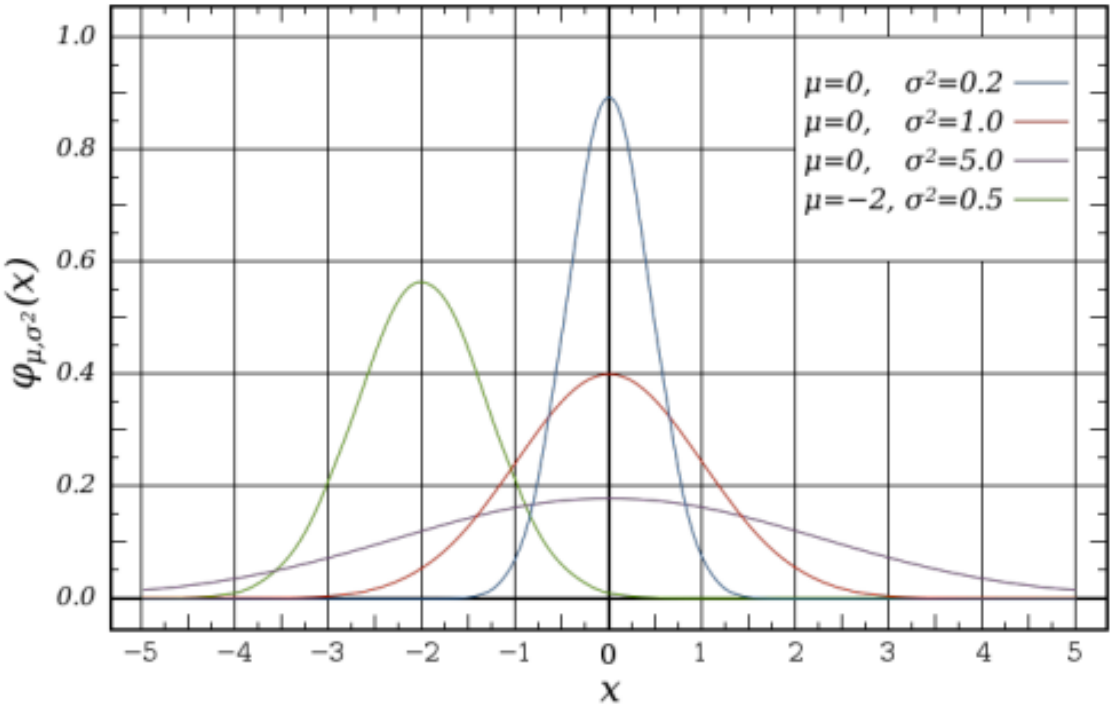
Some important distributions

- **Normal or Gaussian** distribution represents the limit of the Binomial counts when N becomes large

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) = \mathcal{N}(x;\mu,\sigma^2)$$

Normal distribution

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \mathcal{N}(x; \mu, \sigma^2)$$



Normal distribution

In Normal distribution:

- μ is parameter called mean, it also equals the actual mean of X
- σ^2 is parameter called variance, also equals the actual variance of X

These can be verified by calculating directly E[X] and E[(X-μ)^2] using $\mathcal{N}(\mu,\sigma^2)$:

$$E[X] = \int dx x P(x) = \int_{x=-\infty}^{x=\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) dx = \mu$$

Normal distribution

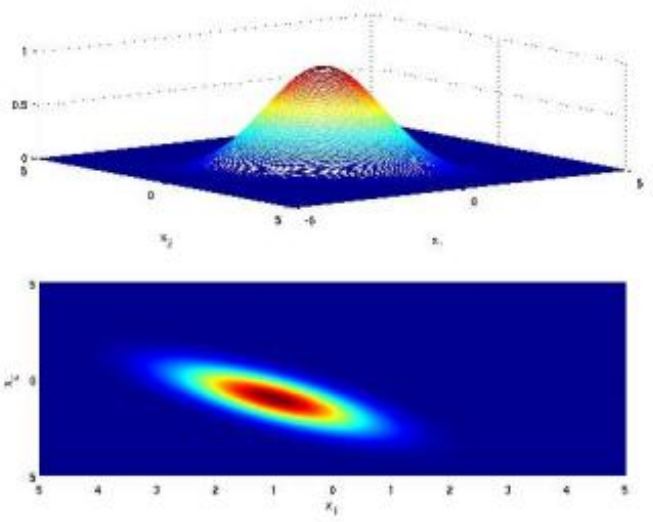
When X is a high-dimensional vector, the Normal distribution is called multivariate:

$$P(x \mid \mu, \Sigma) = \mid 2\pi\Sigma \mid^{-1/2} \exp\left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \right\}$$

Normal distribution

Here $\mu_i = E[X_i]$ is the **vector of the means** and $\sum_{ij}$ is the **covariance matrix**, which controls the shape of the density function

$$P(x \mid \mu, \Sigma) \propto \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$



Normal distribution

The covariance matrix has elements

$$\Sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)] = \mathrm{cov}(X_i, X_j)$$

And can be written in matrix notation as

$$\Sigma = E[(X - \mu)(X - \mu)^T] = E[XX^T] - \mu\mu^T$$

(CHECK THIS AT HOME)

Normal distribution

Some of the important properties of Gaussian:

- All marginals of a Gaussian are Gaussian
- Any conditional of a Gaussian is Gaussian
- The product of two Gaussians is a Gaussian
- Sum of two independent Gaussian r.v. is again a Gaussian

Normal distribution

Gaussian is what is called **limiting distribution** in probability and for that reason appears very often in a variety of settings.

Normal distribution

The **Central Limit Theorem** states that the result of summation of a large number of independent random variables X_i, under very general conditions, is going to be a Gaussian

$$\sum_i X_i \rightarrow \mathcal{N}(\sum E[X_i], \sum var(X_i)),\ n \rightarrow \infty$$

QUESTIONS FOR SELF-CONTROL

- Explain what random variable is. Give examples of r.v.
- Explain what a relationships of random variables is.
- Suppose X is a r.v. How do we understand $X^2 \geq 2X-1$. How can we prove this?
- Suppose X is a r.v. and $Y=(X-1)^2$. What can we say about r.v. Y?
- Define sample space.
- What is the sample space of a r.v. representing the result of rolling double-dice (two dice with the outcomes summed)?
- What is the sample space of rolling a dice two times?
- What is the probability distribution of a r.v.?
- What is mass function of a r.v.?
- Calculate the probability distribution of a r.v. representing the result of rolling double-dice.
- What is the probability density? How is it different from mass function?
- Assume X is in {1,2,3} and P(X) is as follows P(1)=0.3, P(2)=0.4,P(3)=0.5. Is this a valid probability distribution?
- $p(X)=3x^2$ for a continuous r.v. X in [0,1]. Is this a suitable probability distribution?
- Define E[X].
- Calculate E[X] for continuous r.v. X in [0,1] with distribution $p(x)=3x^2$.
- Calculate $E[X^2]$ for continuous r.v. X in [0,1] with distribution $p(x)=3x^2$.
- Define mean, variance, standard deviation, mode an median of a probability distribution.
- What is the mean, variance, standard deviation, mode and median of a Normal probability distribution N(5,4)?

- Define cumulative distribution function (CDF).
- Calculate CDF for r.v. X in [0,1] with distribution $p(x)=3x^2$.
- Define joint probability P(X,Y).
- Define conditional probability P(Y|X).
- Define marginal probability for P(X,Y).
- Define the sum rule and the product/chain rule.
- Write down the Bayes rule. Prove it.
- Define independence of r.v.
- Define conditional independence of r.v.
- Design an example of 3 r.v., X, Y and Z, by specifying P(X,Y,Z), such that X and Y are conditionally independent given Z.
- Prove var(X+Y)=var(X)+var(Y) for independent r.v.
- Prove E[XY]=E[X]E[Y] for independent r.v.
- Write down Bernoulli probability distribution.
- Write down Poisson probability distribution.
- Write down Normal probability distribution.
- Write down multivariate Normal probability density and explain its parameters.
- For discrete probability distribution P(X,Y), X=cabin class {1,2,3,0}, Y=survived {0,1}, in slide 44 calculate: E[X], E[Y], var(X), var(Y). Are X and Y independent?
- For discrete probability distribution above calculate: P(X), P(Y), P(Y|X) and P(X|Y) for each pair of values (X,Y). Check that Bayes rule holds in your results.

- For discrete probability distribution P(X,Y), X=cabin class {1,2,3,0}, Y=survived {0,1}, in slide 44 calculate: E[X], E[Y], var(X), var(Y). Are X and Y independent?
- For discrete probability distribution above calculate: P(X), P(Y), P(Y|X) and P(X|Y) for each pair of values (X,Y). Check that Bayes rule holds in your results.
- For discrete probability distribution above calculate E[X+Y], E[XY], var(X+Y), cov(X,Y) (see slide 60 for definition of covariance cov).
- Repeat the above 3 questions for the probability distributions in slides 42 and 48.
- For continuous probability distribution $p(X,Y)=6x^2y^3$, x and y in [0,1], calculate E[X], E[Y], var(X), var(Y). Are X and Y independent?
- For continuous probability distribution $p(X,Y)=6x^2y^3$, x and y in [0,1], calculate p(X), p(Y), p(Y|X), p(Y|X). Check that Bayes rule holds.
- For continuous probability distribution $p(X,Y)=6x^2y^3$ calculate E[X+Y], E[XY], var(X+Y), cov(X,Y).

ADVANCED

Some important probabilistic inequalities

The following are some simple yet extremely powerful and very important inequalities from probability, which play significant role in advanced statistical learning.

Interested students may further inquire about ***the theory of large deviations*** related to this topic.

Some important probabilistic inequalities

Jensen's inequality allows one to convert the expected value of a convex function of a random variable into a function of the expected value of that r.v. Jensen's inequality exploits the basic property of the convex functions
$f(\sum \alpha_i x_i) \leq \sum \alpha_i f(x_i)$, for any $\sum \alpha_i = 1$.

Jensen's inequality is extremely widely used.

Some important probabilistic inequalities

Jensen's inequality:

$$E[f(X)] \geq f(E[X])$$

Derivation:

$$E[f(X)] = \sum_x f(x)P(x) = \sum_x P(x)f(x) \ / by\ convexity/$$

$$\geq f(\sum_x P(x)x) = f(E[X])$$

Some important probabilistic inequalities

Markov inequality is one of the most important yet unbelievably simple inequalities in probability. It allows one to bound a random variable by knowing practically nothing about it.

Markov inequality: For any non-negative r.v. $X \geq 0$ and any number $a \geq 0$:

$$\Pr\{X \geq a\} \leq \frac{E[X]}{a}$$

Some important probabilistic inequalities

Proof: Introduce a random variable I_a=1 if X$\geq a$ and 0 otherwise (**the indicator r.v.**). Note that E[I_a]=$\sum_{x\geq a}$P(x)= Pr{X$\geq a$}. At the same time, $aI_a \leq$X in the random inequality sense (for all realizations). Then, aPr{X$\geq a$}=E[aI_a]$\leq$E[X], which gives the statement.

Some important probabilistic inequalities

Example:

Assuming on average it takes me 20 minutes to arrive from F.Altay to IEU, the probability that it will take me more than 60 minutes to come is $\Pr\{X \geq 60 \text{ min}\} \leq 20/60 = 1/3$. I don't need to know anything about X other than its mean value.

Some important probabilistic inequalities

Chebyshev inequality is also known as the **Law or Large Numbers** and states that an average of large number of independent r.v. Xi will tend to the mean of E[Xi] via a particularly ingenious use of the Markov's inequality:

$$\Pr\{|X-\mu|\geq\varepsilon\}=\Pr\{|X-\mu|^2\geq\varepsilon^2\}\leq\frac{E[(X-\mu)^2]}{\varepsilon^2}=\frac{\mathrm{var}(X)}{\varepsilon^2}$$

$$\Pr\{|X-\mu|\geq\varepsilon\}\leq\frac{\mathrm{var}(X)}{\varepsilon^2}$$

Some important probabilistic inequalities

Here $\mu=E[X]$ is the expectation value of X.

By substituting $X=S_n=1/n\sum_i X_i$ (an average of n independent r.v. X_i), we can see that the probability of S_n deviating from $1/n\sum_i E[X_i]$ by *any finite amount* ε tends to zero at least as 1/n, which is the **Law of Large Numbers** :

$$\Pr\{|S_n - \frac{1}{n}\sum_{i=1}^{n} E[X_i]| \geq \varepsilon\} \leq \frac{\text{ave}(\text{var}(X_i))}{n\varepsilon^2}$$

Some important probabilistic inequalities

A stronger version of the Law of Large Numbers is known as the **Central Limit Theorem**, which states that not only S_n will tend to the arithmetic mean of Xi in probability as $n \to \infty$, but also will have a very specific distribution, which is the **Normal distribution** with the mean equal to the average of $E[X_i]$ and the variance equal to the average of $var(X_i)$ and divided by n

Some important probabilistic inequalities

Central Limit Theorem for independent r.v. Xi and $S_n=1/n\sum_{i=1..n}X_i$:

$$P(S_n) \to N(\frac{1}{n}\sum_{i=1}^{n} E[X_i], \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(X_i))$$

If X_i are independent copies of the same X, $X_i=X$, then we get the more commonly known form of CLT:

$$P(S_n) \to N\left(E[X], \frac{\mathrm{var}(X)}{n} \right)$$

Some important probabilistic inequalities

Chernoff inequality or bound is an extension of Markov's inequality to any X:

$$\Pr\{X \geq a\} = \Pr\{e^{tX} \geq e^{ta}\} \leq \frac{E[e^{tX}]}{e^{ta}}, t \geq 0$$

Some important probabilistic inequalities

Hoeffding bound is a related and more advanced version of Chernoff inequality, which is formulated as follows:

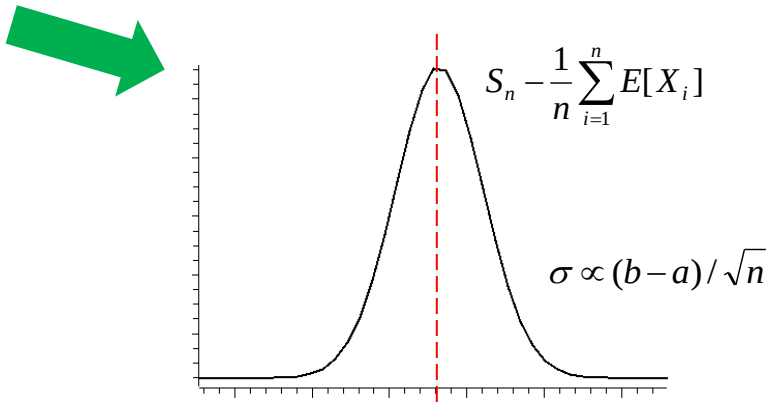
Let $X_1,\ldots,X_n$ be independent bounded r.v. $a \leq X_i \leq b$. Then:

$$\Pr\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - E[X_i]) \geq t\right\} \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right), \text{ for any } t \geq 0.$$

Some important probabilistic inequalities

Hoeffding bound is a very strong (Gaussian in fact) bound with respect to the deviation of Sn from the mean of expectations, and is obtained by taking advantage of the **bounded** nature of the summed r.v.

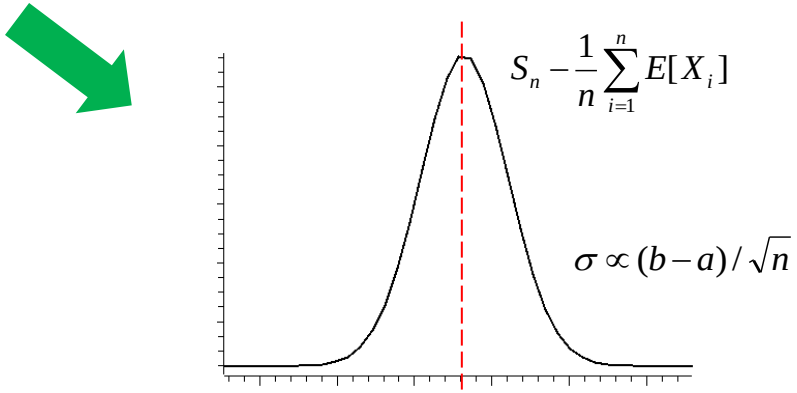
$$\Pr\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - E[X_i]) \geq t\right\} \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$



Some important probabilistic inequalities

Hoeffding bound is also very general as only independence and boundness of Xi is required.

$$\Pr\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - E[X_i]) \geq t\right\} \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$



Bayesian estimation

Bayesian estimation is a rigorous recipe in probability theory for accumulating information. Bayesian estimation is extremely widely used in modern data processing, statistics, and ML.

At the basis of Bayesian estimation is the Bayes theorem:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

Bayesian estimation

Consider the problem: suppose we know that observations X come from a process described by parametric generative probability density $P_\theta(X)$. Suppose also that we have collected a series of observations Xi. What is the value for θ that we must deduce from such observations?

Bayesian estimation

The answer to that question is that, in fact, there isn't any. Our knowledge about θ is described by the probability distribution $P(\theta|\{Xi\})$, which in turn is expressed by the Bayes theorem

$$P(\theta|\{X_i\}) = \frac{P(\{X_i\}|\theta)P(\theta)}{P(\{X_i\})}$$

Bayesian estimation

Here, P(θ) is known as the **prior probability distribution** and expresses our belief about θ before seeing any data. P(θ|{Xi}) then expresses our knowledge about θ considering the obtained data, and is called **posterior probability**.

$$P(\theta|\{X_i\}) = \frac{P(\{X_i\}|\theta)P(\theta)}{P(\{X_i\})}$$

Bayesian estimation

One completely reasonable choice P(θ)=const is called **uniform** or **non-informative prior**, which expresses the fact that we have no preferences for one θ or another before any data. Then our knowledge of θ comes only from data and can be expressed as

$$P(\theta|\{X_i\}) \propto P(\{X_i\}|\theta)$$

Bayesian estimation

However, often stronger priors that significantly constrain the uncertainty about P(θ|{Xi}) can be found, thus dramatically improving our ability to estimate θ with much more limited data:

$$P(\theta \mid \{X_i\}) = \frac{P(\{X_i\} \mid \theta) P(\theta)}{P(\{X_i\})}$$

Bayesian estimation

If data is acquired gradually, by denoting $P(\theta|\{Xi,i=1..n\})=P_n(\theta)$ and assuming that observations $P(Xi|\theta)$ are mutually independent, we can obtain the following Bayes rule for incremental learning (of θ):

$$P_{n+1}(\theta) = \frac{P(X_{n+1}|\theta)P_n(\theta)}{P(X_{n+1})}$$

Bayesian estimation

Maximal a-posteriory estimation, or **MAP**, selects a single value for θ on the basis of maximization of P(θ|{Xi}). MAP dictates that the best estimate for θ given data {Xi} is such that has the highest posterior probability P(θ|{Xi})

$$\theta_{MAP} = \arg\max P(\theta \mid \{X_i\})$$

Bayesian estimation

Note that MAP cannot be done incrementally, as the incremental Bayes rule requires knowledge of the entire previous distribution $P_n(\theta)$. However, approximate schemes can be used such as approximating $P_n(\theta)$ with a Gaussian, etc.

$$P_{n+1}(\theta) = \frac{P(X_{n+1} \mid \theta) P_n(\theta)}{P(X_{n+1})}$$

$$\theta_{MAP} = \arg\max P(\theta \mid \{X_i\})$$

Bayesian estimation

The quantities in the denominator such as P({Xi}) are the probabilities of observing {Xi} given any generative process θ, and are the marginalization of P({Xi},θ) over θ. Such quantities are notoriously difficult to compute.

$$P_{n+1}(\theta) = \frac{P(X_{n+1} \mid \theta) P_n(\theta)}{P(X_{n+1})}$$

$$P(\theta \mid \{X_i\}) = \frac{P(\{X_i\} \mid \theta) P(\theta)}{P(\{X_i\})}$$

Bayesian estimation

Formally, we can write P({Xi},θ) = P({Xi}|θ)P(θ) and then Σ_θP({Xi},θ) = Σ_θ P({Xi}|θ)P(θ), resulting in the following Bayes rule:

$$P_{n+1}(\theta) = \frac{P(X_{n+1}\,|\,\theta)P(\theta)}{\sum_{\theta'} P(X_{n+1}\,|\,\theta')P(\theta')}$$

$$P(\theta\,|\,\{X_i\}) = \frac{P(\{X_i\}\,|\,\theta)P(\theta)}{\sum_{\theta'} P(\{X_i\}\,|\,\theta')P(\theta')}$$

Bayesian estimation

These denominator sums are known as normalization constants, since they normalize P(θ) to be 1 after new observations were incorporated via the Bayes theorem:

$$P_{n+1}(\theta) = \frac{P(X_{n+1}|\theta)P(\theta)}{\sum_{\theta'} P(X_{n+1}|\theta')P(\theta')} = \frac{P(X_{n+1}|\theta)P(\theta)}{Z}$$

$$P(\theta|\{X_i\}) = \frac{P(\{X_i\}|\theta)P(\theta)}{\sum_{\theta'} P(\{X_i\}|\theta')P(\theta')} = \frac{P(\{X_i\}|\theta)P(\theta)}{Z}$$

Bayesian estimation

Since normalization constants do not depend on θ, they are not relevant in the context of MAP estimation, and can be discarded therefore allowing us to estimate θ in MAP without having to worry about difficult to compute integrals Z

$$\theta_{MAP} = \arg\max P(\theta \mid \{X_i\}) = \arg\max P(\{X_i\} \mid \theta) P(\theta)$$

Bayesian estimation

While previous discussion was inclusive of either dependent or independent r.v. Xi, if Xi are known to be independent from each other, than

$$P(\{X_i\}|\theta)P(\theta) = P(\theta)\prod_i P(X_i|\theta)$$

and an advantageous concept for MAP inference is that of log-likelihood:

$$\begin{aligned}\theta_{MAP} &= \arg\max(\log\{P(\{X_i\}|\theta)P(\theta)\}) \\ &= \arg\max(\sum_i \log P(X_i|\theta) + \log P(\theta)) \\ &= \arg\max(\text{loglik}(\theta) + \log P(\theta))\end{aligned}$$

Bayesian estimation

If a uniform prior is assumed, P(θ)=const, such prior drops out from MAP estimation leading to what is also known as **Maximum Likelihood Estimation** rule, or **MLE**

$$\theta_{MLE} = \arg\max \sum_i \log P(X_i \mid \theta) = \arg\max \text{loglik}(\theta)$$

Bayesian estimation

In MLE, the best choice for θ given observations {Xi} is said to be such that maximizes the probability of such observations to be observed

$$\theta_{MLE} = \arg\max \log P(\{X_i\} | \theta)$$

Bayesian estimation

Bayesian estimation differs from MLE in the use of additional priors to allow for a faster estimation of θ, that is, with less data. The prior, therefore, takes on the role of substituting some of the information about θ that would otherwise normally have to be obtained from the observations Xi.

$$\theta_{MAP} = \arg\max \log P(\{X_i\} | \theta) P(\theta)$$

$$\theta_{MLE} = \arg\max \log P(\{X_i\} | \theta)$$