Lecture 1:
Basics of Probability

(Luise-Vitetta, Chapter 8)

Ilenia Tinnirello

Why probability in data science?

- ➔ Data ***acquisition*** is noisy
  - ⇨ Sampling/quantization
  - ⇨ external factors: If you record your voice saying 'machine learning' you will never get the same function!
    - → The function is deterministic, once recordered, but it is not predictable as f(machine learning)!
- ➔ Data ***analysis*** often works on predictions.
  - ⇨ We live in a world of untecertainty!
    - → It rains tomorrow? How many cm?

Probability is the core mathematical tool for working with noisy data and uncertain events

- ⇨ Some concepts are not intuitive and we need some basic theory!

Counting Outcomes: some examples (intuitive and not)

Three important remarks

- ➔ **For independent events, the hystory does not matter!!**
  - ⇨ After 10 heads, the probability of tossing a coin and getting another head is still ½!
- ➔ **For combinatorial analysis of objects of the same class, each object has to be considered singularly!**
  - ⇨ What it the probability fo extracting two consecutive white balls from a set of white and black balls in a box?
- ➔ **Information matters!**
  - ⇨ A priori probability can be updated, once partial information is disclosed!

Head or Tail?

- **Consider one unbiased coin**
  - In one flip, ½ head, ½ tail
  - In two flips?
    - HH, HT, TH, TT each with equal probability (1/4)
  - Assume that we already did 10 flips with the outcome HHHHHHHHHH
    - What is the probability to have now head or tail at the next flip?

Head or Tail?

- ➔ **Two soccer teams are equally good: they have the same probabiltiy to win a match**
- ➔ **What is the most likely length for reaching 4 successes?**
  - ⇨ Flip a coin until 4 heads or 4 tails are reached.
  - ⇨ Is it more likely to have 6 or 7 flips?
- ➔ **To reach 4 successes, after 5 flips, we need to have 3H, 2T or 3T, 2H**
  - ⇨ ½ it ends 4 to 2 (6 flips)
  - ⇨ ½ it ends 4 to 3 (7 flips)

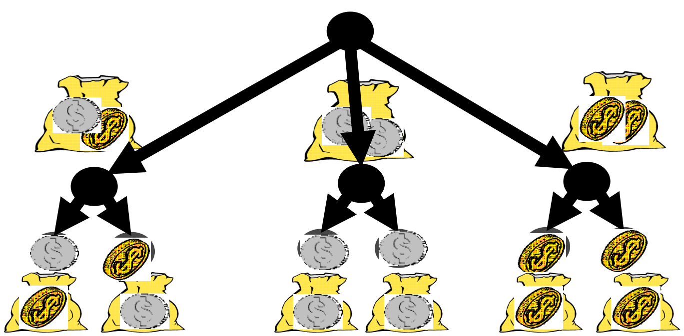
Silver or Gold

- One bag has two silver coins, another has two gold coins, and the third has one of each
- One bag is selected at random. One coin from it is selected at random. It turns out to be gold
- What is the probability that the other coin is gold?
  - Be careful to ‘ordered’ events!!!!

Analysis of Outcomes

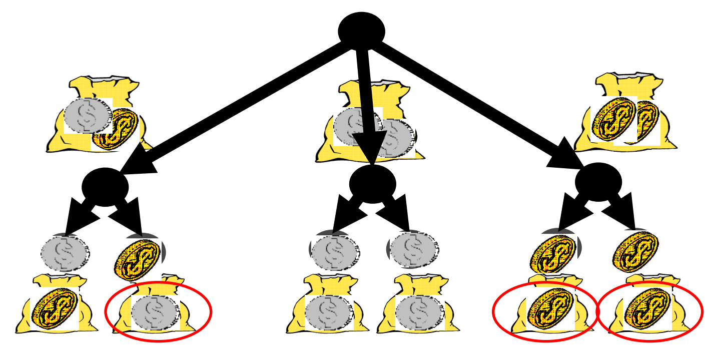
A priori:

- ➔ 3 choices of bag
- ➔ 2 ways to order bag contents
- ➔ 6 equally likely **outcomes**



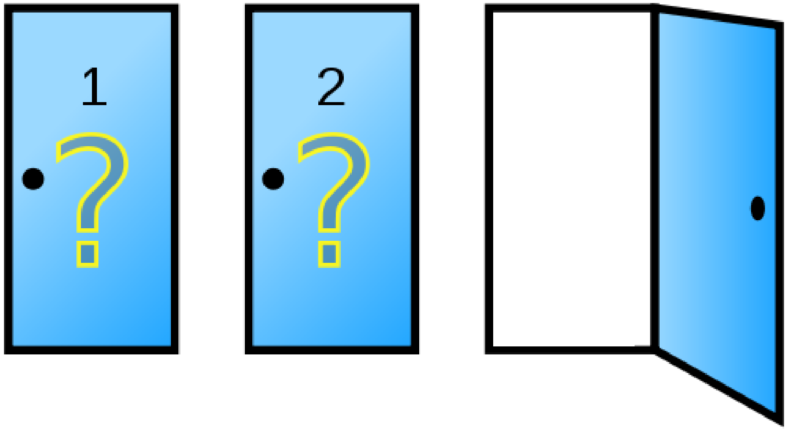
After we see a gold coin:

- ➔ only **3 outcomes** are possible
- ➔ two of the remaining three events are gold coins
  - ⇨ 2/3 probability2 ways to order bag contents



Monty Hall Problem

- **Announcer hides prize behind one of 3 doors at random**
- **You select some door**
- **Announcer opens one of others with no prize**
- **You can decide to keep or switch**
  - What to do?



Analysis of Outcomes

- ➔ Sample space = {prize behind door 1, prize behind door 2, prize behind door 3}
- ➔ Each has probability 1/3

Staying:

We win, if we chose the correct door

Probability = 1/3

Switching:

We win, if we chose the incorrect door

Probability = 2/3

Trick: "After one door is opened, others are equally likely..."

But his action is not independent of yours!

Information matters.. we will see more later

Imagine a similar problem with 100 doors!

Basic Theory

Experiments

- **Experiment :** any process or procedure for which more than one outcome is possible
- **Sample Space:** The sample space *S* of an experiment is a set consisting of all of the possible experimental outcomes

- **Example 1:** a manager supervises the operation of three power plants, at any given time, each of the three plants can be classified as either generating electricity (1) or being idle (0).

$$S = \{(0,0,0)\ (0,0,1)\ (0,1,0)\ (0,1,1)\ (1,0,0)\ (1,0,1)\ (1,1,0)\ (1,1,1)\}$$

- **Example 2:** the roll of a die has S={1, 2, 3, 4, 5, 6}
- **Example 3:** the roll of two dice has 36 possible outcomes S={(1, 1), (1,2), (1, 3), … (6, 1), (6, 2), …(6,6)}

Events and Complements

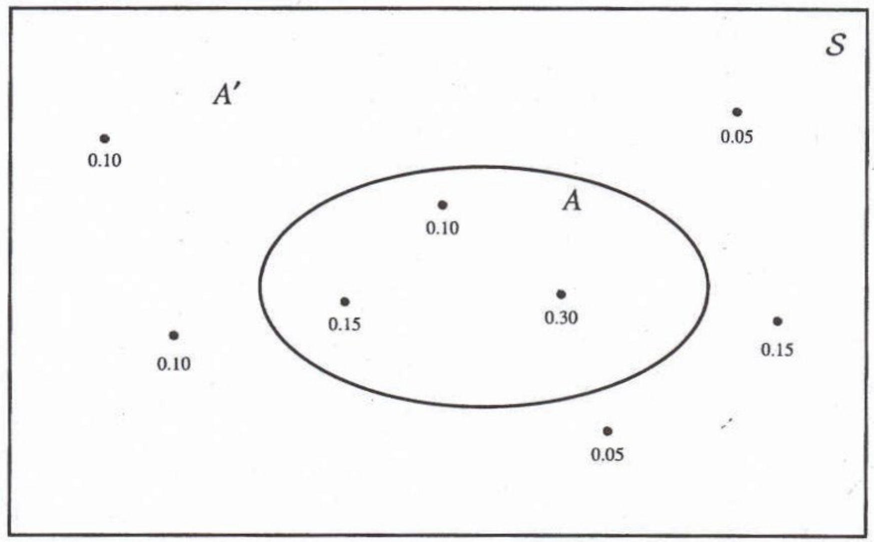
- **An event A is a subset of the sample space S. It collects outcomes of particular interest.**
  - The probability of an event is obtained by summing the probabilities of the outcomes contained within the event A
- **An event is said to occur if one of the outcomes contained within the event occurs.**
- **Events that consist of an individual outcome are sometimes referred to as elementary events or simple events**
- **If A and B are events, then A∪B is an event too**
- **the complement of event *A*, is the event consisting of everything in the sample space *S* that is not contained within the event *A*.**

Axiomatic Definition

- **Probability Space: (Sample Space, Event Class, Pr law)**
- **Axioms:**
  - $Pr(A)>0;$
  - $Pr(S)=1;$
  - $A \cap B=\varnothing \rightarrow Pr(A \cup B)=Pr(A)+Pr(B)$
- **It follows:**
  - $Pr(S-A)=1-Pr(A)$
  - $Pr(A)<=1$
  - $Pr(A \cup B)=Pr(A)+Pr(B)-Pr(B \cap A)$
- **$Pr(B \cap A)$ is called joint probability**
- **$Pr(B/A)$ is called conditioned probability**
  - $Pr(B/A)=Pr(A \cap B)/Pr(B)$

Events and Complements

→ **A sample space S consists of eight outcomes with a probability value.**



$$P(A) = 0.10 + 0.15 + 0.30 = 0.55$$

$$P(A') = 0.10 + 0.05 + 0.05 + 0.15 + 0.10 = 0.45$$

$$\text{Notice that } P(A) + P(A') = 1.$$

Events and Complements

➔ **GAMES OF CHANCE**

- even = { an *even* score is recorded on the roll of a die }
= { 2,4,6 }

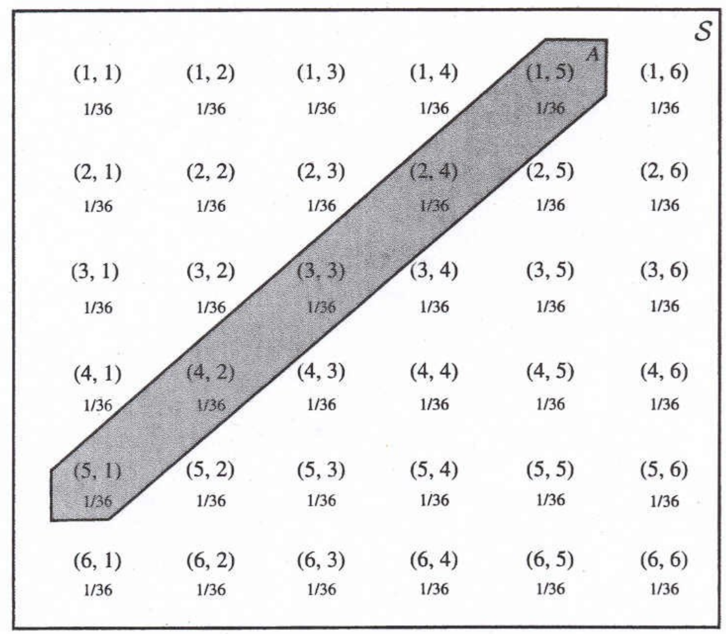
$$P(\text{even}) = P(2) + P(4) + P(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

- A = { the sum of the scores of two dice is equal to 6 }
= { (1,5), (2,4), (3,3), (4,2), (5,1) }

$$P(A) = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{5}{36}$$

A sum of 6 will be obtained with two fair dice roughly 5 times out of 36 on average, that is, on about 14% of the throws.

- **B = { at least one of the two dice records a 6 } = ????**



Examples: Intersection of Events

FIGURE 1.26
Events A and B

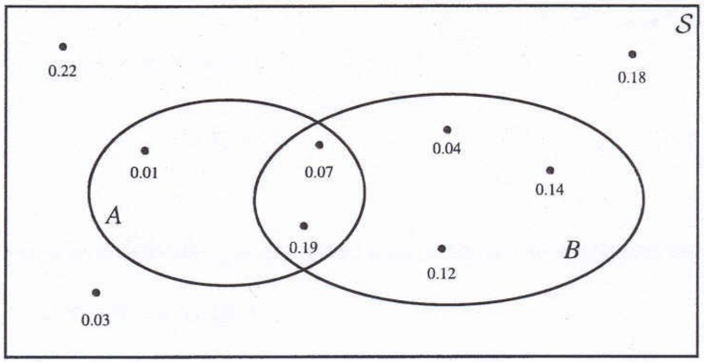
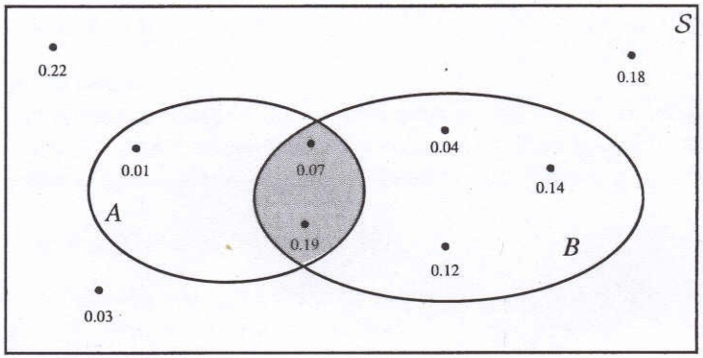


FIGURE 1.27
The event $A \cap B$



$$P(A) = 0.01 + 0.07 + 0.19 = 0.27$$

$$P(B) = 0.07 + 0.19 + 0.04 + 0.14 + 0.12 = 0.56$$

$$P(A \cap B) = 0.07 + 0.19 = 0.26$$

Example: Intersection of Events

FIGURE 1.28
The event $A' \cap B$

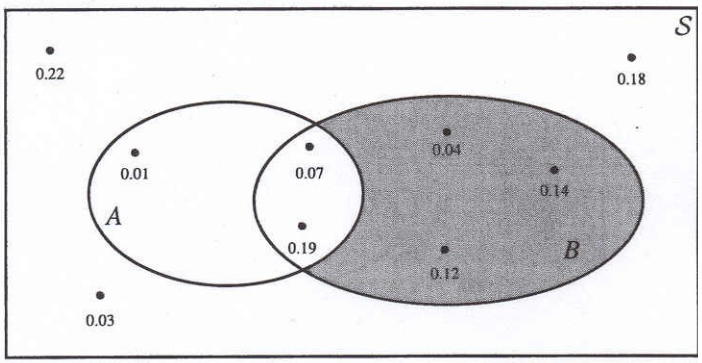
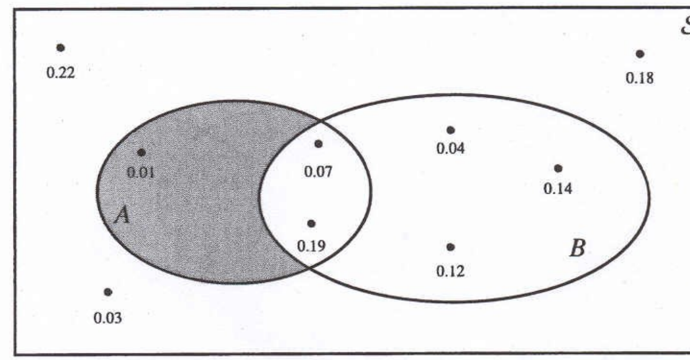


FIGURE 1.29
The event $A \cap B'$



$$P(A' \cap B) = 0.04 + 0.14 + 0.12 = 0.30$$

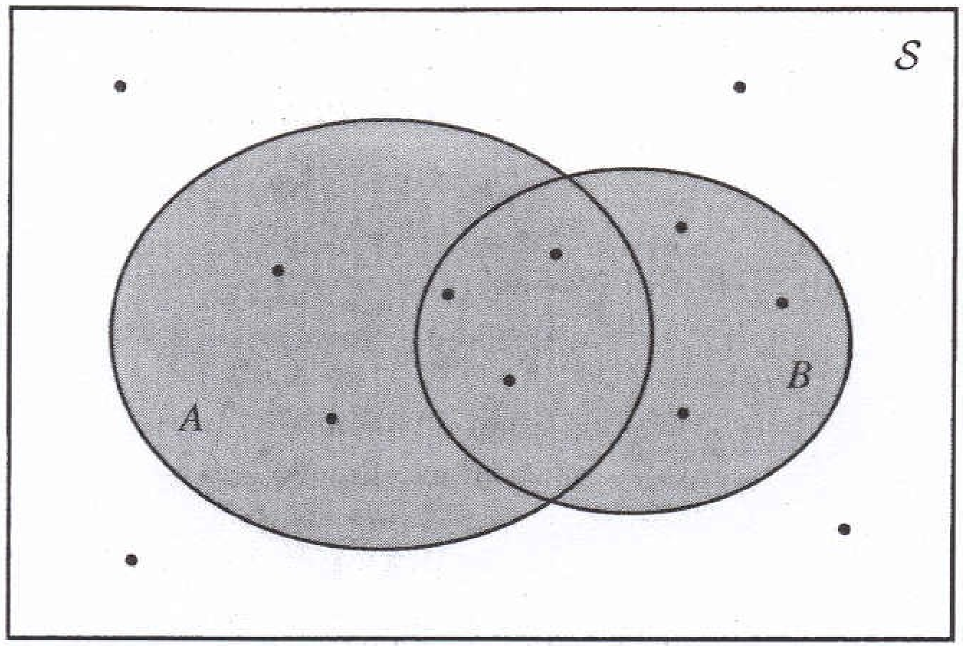
$$P(A \cap B') = 0.01$$

$$P(A \cap B) + P(A \cap B') = 0.26 + 0.01 = 0.27 = P(A)$$

$$P(A \cap B) + P(A' \cap B) = 0.26 + 0.30 = 0.56 = P(B)$$

Example: Union of Events

FIGURE 1.32 ●
The event $A \cup B$



➔ **It includes outcomes in A, outcomes in B, outcomes in A and B.**

Example: Conditional Probabilities

➔ **GAMES OF CHANCE**

- A fair die is rolled.

$$P(6) = \frac{1}{6}$$

$$P(6 \mid even) = \frac{P(6 \cap even)}{P(even)} = \frac{P(6)}{P(even)}$$

$$= \frac{P(6)}{P(2) + P(4) + P(6)} = \frac{1/6}{1/6 + 1/6 + 1/6} = \frac{1}{3}$$

- A red die and a blue die are thrown.

A **= { the red die scores a 6 }**

B **= { at least one 6 is obtained on the two dice }**

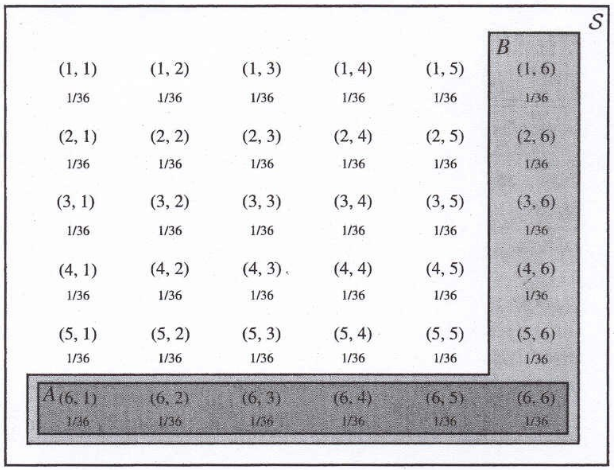
$$P(A) = \frac{6}{36} = \frac{1}{6} \text{ and } P(B) = \frac{11}{36}$$

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

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

$$= \frac{1/6}{11/36} = \frac{6}{11}$$

FIGURE 1.60 •
$P(A|B) = P(A \cap B)/P(B)$



Bayes Theorem

- Event probability conditioned to the occurrence of other certain events
  - The event space is no more the total space, but it becomes the conditioning event
- $Pr(A/B)=Pr(A \cap B)/Pr(B)$
  - $Pr(A \cap B)=Pr(A/B)\ Pr(B) = Pr(B/A)\ Pr(A)$
  - $Pr(A/B)=Pr(B/A)\ Pr(A)/Pr(B)$

Bayes Theorem Relevance

- ➔ A mammograms detects cancer in 80% of the cases, if there is
- ➔ A mammograms detects a false positive in 9.6% of the cases, if there is no cancer
  - ⇨ Test positive: 80% YES, 9.6% NO
  - ⇨ Test negative: 20% YES, 90.4% NO
- ➔ Cancer incidence is about 1% of woman population.
- ➔ What is the percentage to have a cancer is the result is positive?
  - ⇨ Possible positive events:
    - → 80% of 1% of population: 0.008;
    - → 9.6% of 99% of population: 0.095;
  - ⇨ Conditioned probability:
    - → Real positive results / total events = 0.008/(0.008+0.095)=0.0776

Bayes Theorem and Monty Hall

- Why Bayes and the Monty Hall Problem?
  - At first, 3 equal doors A, B, C with Pwinning probability equal to 1/3
  - Assume we pick door A and the opened door is C.
    - P(C open)=P(A car) * ½ + [1-P(A car)]*(P(B car/A no car))*1 + P(C car/A no car))* 0) =1/2*1/3 + 2/3*(1/2+0)=1/2.
- What we **learn** on door A, after one door is opened, given that the opened door will be always an empty door?
  - Nothing! The selected door does not depend in any case from door A.
  - P(A car/C open)=P(C open/A car)*P(A car)/P(C open)=P(A car)=1/3
- What we **learn** on door B?
  - A lot! The selected door depends on node B state.
  - P(B car/C open)=P(C open/B car) P(B car)/P(C open)=1*1/3/0.5=2/3
- Can we generalize to the 100 doors case?

Monty Hall Generalization

1. Let A be the first selected door, X be the final door left closed and let C1, C2, .. C98 the other doors sequentially opened. At each step:
   - ⇨ P(A car/C1 opened) = P(C1 opened/ A car) P(A car)/P(C1 opened)=P(A car)=1/100
   - ⇨ P(X car/C1 opened) = P(C1 opened/ X car) P(X car)/P(C1 opened)=1/98 * 1/100 * 99=99/98 * 1/100 -> I am improving over 1/100!
2. Now we open C2:
   - ⇨ P(X car/C1, C2 open)=P(C1, C2 open/X car) P(Xcar) / P(C1, C2 open)=
   - ⇨ P(C2 open/C1 open, X car)*P(C1 open, X car)/[P(C2 open/C1 open)*P(C1 open)]=1/97 * 1/98*1/100 / (1/98 * 1/99) = 99/97 * 1/100
3. Now we open C3, C4, ... C98
   - ⇨ P(X car/C1, C2, … C98 open) = 99/100!

Other Pr definitions

- **Frequency analysis**
  - Limit for number of experiments growing to infinity of event occurrence
  - Kolmogorov axioms verified

Composing Experiments

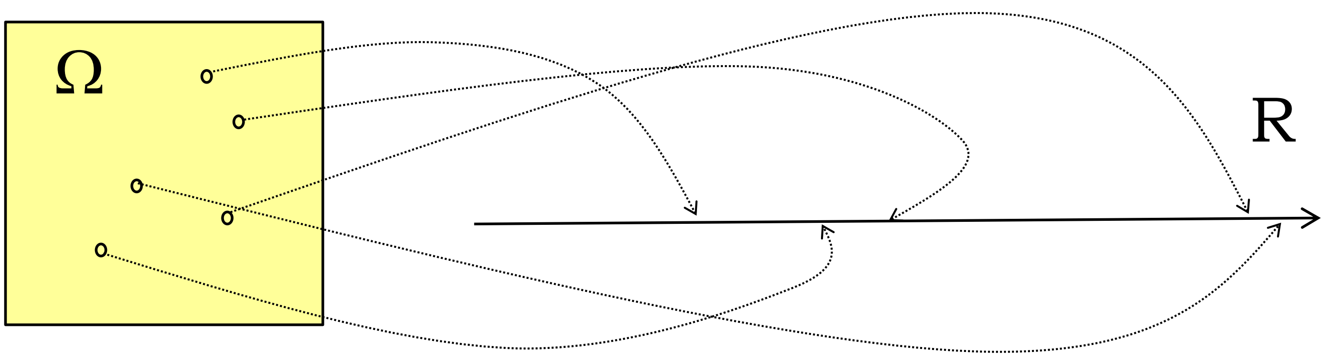
- **Given two experiments with sample space S1 and S2, an ordered couple of outcomes from each one is a composed experiment**
- **If experiments are independent**
  - $Pr(A1 \times A2) = Pr(A1) Pr(A2)$
- **..otherwise it is not possible to have the probability law of the composed experiment**

- **Bernoulli formula:**
  - Probability of n identical experiments with outcomes 0/1

Random Variables

Definition

- **We can associate a numerical value to an event**
  - E.g. 0 to an head tossing, 1 to a tail tossing
- **Formally:**
  - Let (Ω, S, Pr) a probability space with sample space Ω, event class S and probability law Pr
  - Let $X(\omega) \in R$ a real number associated to each experiment outcome ω
  - $X(\Omega)$, i.e. the set of all the outcome images in R is a random variable if:
    - $\forall a \in R$, the set $\{\omega : X(\omega) < a\}$ is an event

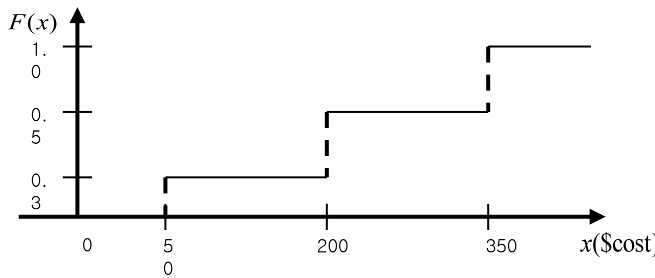


Probability Distribution Function

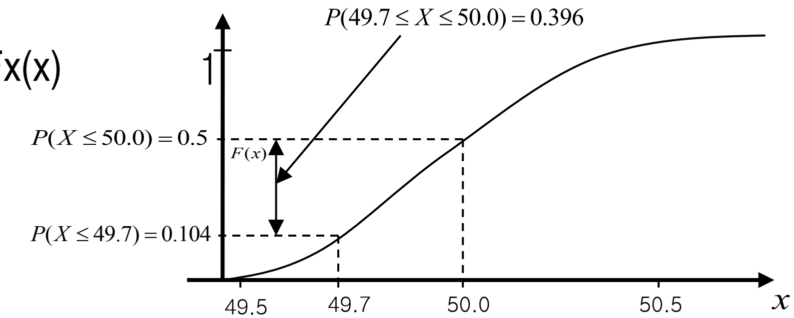
- **From the definition or random variable, it follows that the set of ω values for which $X(\omega)<a$ is an event**
  - $\{\omega: a<X(\omega)<=b\}$ is also an event with its occurrence probability
  - We can indicate the event probability as $Pr\{a<X<=b\}$
- **Distribution function:**
  - $Fx(x)=Pr\{X<=x\}$
- **Properties:**
  - $0<=Fx(x)<=1$
  - $\lim_{x\to\infty} Fx(x)=1$
  - $\lim_{x\to-\infty} Fx(x)=0$
  - $x2>x1,\ Fx(x2)>=Fx(x1)$
  - $\lim_{h\to o+} Fx(x+h)=Fx(x)$
  - If x^* is a point in which the function is not continous, $Pr\{X=x^*\}=Fx(x^*+)-Fx(x^*-)$
  - $Pr\{a<X<=b\}=Fx(b)-Fx(a)$

Discrete, Continous and Mixed Variables

- ➔ **Depending on Fx, a random variable can be discrete, continous or mixed**
  - ⇨ X may assume only a finit set of values
    - → The probability to have these values is called probability mass function



  - ⇨ X may assume continous values in R with Fx(x) continous function



  - ⇨ X may assume continous values, but Fx(x) with discontinuities

Probability density function

- **Alternative description of random variables**
  - fx(x)=D[Fx(X)]
  - From which: $\int_{-\infty}^{x} fx(t)dt$
- **Properties:**
  - fx(x)>=0
  - Pr{a<X<=b}=Fx(b)-Fx(a)= $\int_{a}^{b} fx(t)dt$
  - $\int_{-\infty}^{\infty} fx(t)dt = 1$
- **Examples:**
  - Uniform distribution
  - Discrete variables
  - Exponential variables (often used for device lifetime)

Expectations

➔ **Expectation of a discrete random variable with p.m.f**

$$P(X = x_i) = p_i \longrightarrow E(X) = \sum_i p_i x_i$$

➔ **Expectation of a continuous random variable with p.d.f** $f(x)$

$$E(X) = \int_{\text{state space}} x f(x) dx$$

➔ **The expected value of a random variable is also called the mean of the random variable**

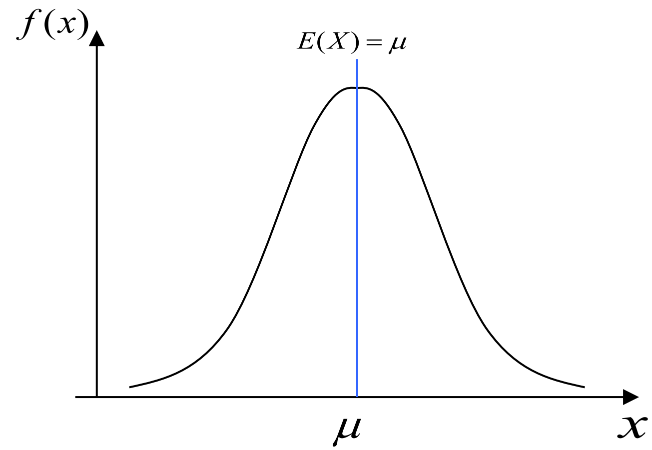
Expectations of Continuous Random Variables

➔Symmetric Random Variables

⇨If x has a p.d.f $f(x)$ that is symmetric about a point μ

so that $f(\mu + x) = f(\mu - x)$

⇨Then, $E(X) = \mu$ (why?)

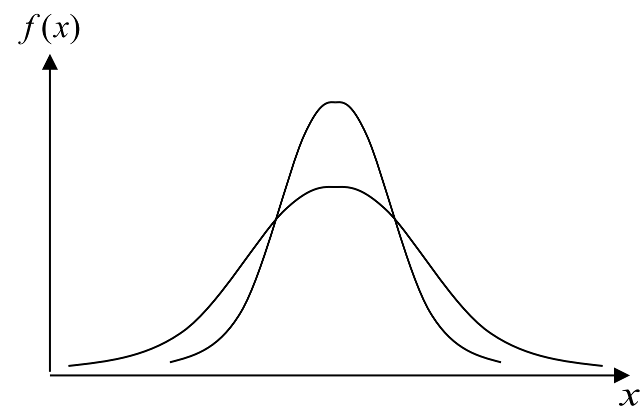


⇨So that the expectation of the random variable is equal to the **point of symmetry**

Variance

- **Variance(σ^2)**
  - A positive quantity that measures the spread of the distribution of the random variable about its mean value
  - Larger values of the variance indicate that the distribution is more spread out
  - Definition: $\mathrm{Var}(X) = E((X - E(X))^2)$

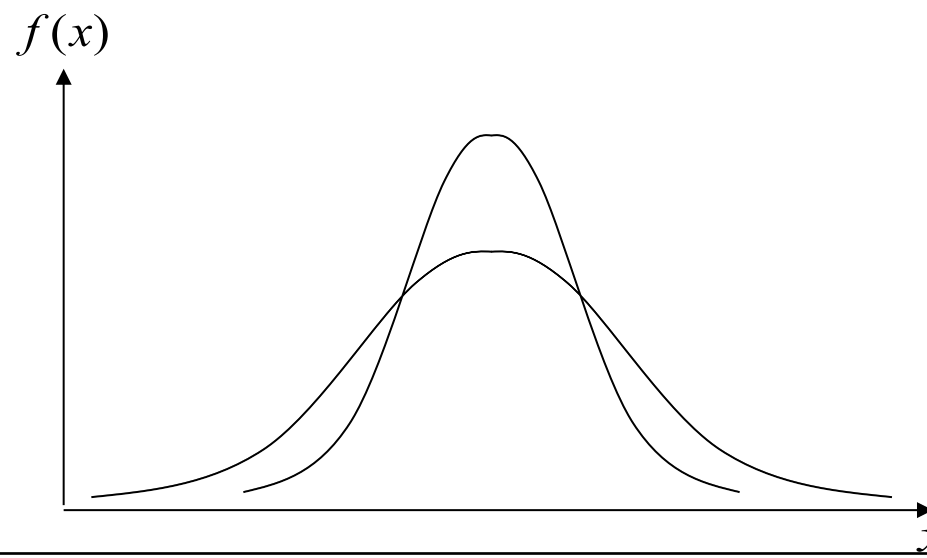
    $$= E(X^2) - (E(X))^2$$
- **Standard Deviation**
  - The positive square root of the variance
  - Denoted by σ
- **Example: gaussian variable**
  - Z=sum_i X_i i.i.d., fz(z) is gaussian!



Interpretation of Variance

$$
\begin{aligned}
\mathrm{Var}(X) &= E((X-E(X))^2) \\
&= E(X^2 - 2XE(X) + (E(X))^2) \\
&= E(X^2) - 2E(X)E(X) + (E(X))^2 \\
&= E(X^2) - (E(X))^2
\end{aligned}
$$

Two distribution with identical mean values but different variances



Jointly Distributed Random Variables

➔ **Joint Probability Distributions**

⇨ Discrete $P(X = x_i, Y = y_j) = p_{ij} \geq 0$

$$\text{satisfying} \quad \sum_i \sum_j p_{ij} = 1$$

⇨ Continuous $f(x,y) \geq 0$ satisfying $\iint_{\text{state space}} f(x,y)dxdx = 1$

➔ **Joint Cumulative Distribution**

⇨ Discrete $F(x,y) = P(X \leq x_i, Y \leq y_j)$

$$F(x,y) = \sum_{i:x_i \leq x} \sum_{j:y_j \leq y} p_{ij}$$

⇨ Continuous

$$F(x,y) = \int_{w=-\infty}^{x} \int_{z=-\infty}^{y} f(w,z)dzdw$$

Marginal Probability Distributions

➔ Marginal probability distribution

- ⇨ Obtained by summing or integrating the joint probability distribution over the values of the other random variable
- ⇨ Discrete $$P(X=i\ )=p_{i+}=\sum_{j}p_{ij}$$
- ⇨ Continuous $$f_X(x)=\int_{-\infty}^{\infty}f(x,y)dy$$

Conditional Probability Distributions

➔ Conditional probability distributions

- ⇨ The probabilistic properties of the random variable X under the knowledge provided by the value of Y
- ⇨ Discrete

$$p_{i|j} = P(X = i \mid Y = j) = \frac{P(X = i, Y = j)}{P(Y = j)} = \frac{p_{ij}}{p_{+j}}$$

- ⇨ Continuous

$$f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)}$$

- ⇨ The conditional probability distribution is a **probability distribution.**

Independence and Covariance

- **Two random variables X and Y are said to be independent if**
  - Discrete

$$p_{ij} = p_{i+}p_{+j} \quad \text{for all values } i \text{ of } X \text{ and } j \text{ of } Y$$

  - Continuous

$$f(x, y) = f_X(x) f_Y(y) \quad \textit{for all } x \textit{ and } y$$

Independence and Covariance

➔ Covariance

$$\mathrm{Cov}(X,Y) = E((X - E(X))(Y - E(Y)))$$
$$= E(XY) - E(X)E(Y)$$

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E((X - E(X))(Y - E(Y))) \\ &= E(XY - XE(Y) - E(X)Y + E(X)E(Y)) \\ &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y)\end{aligned}$$

- ⇨ May take any positive or negative numbers.
- ⇨ Independent random variables have a covariance of zero

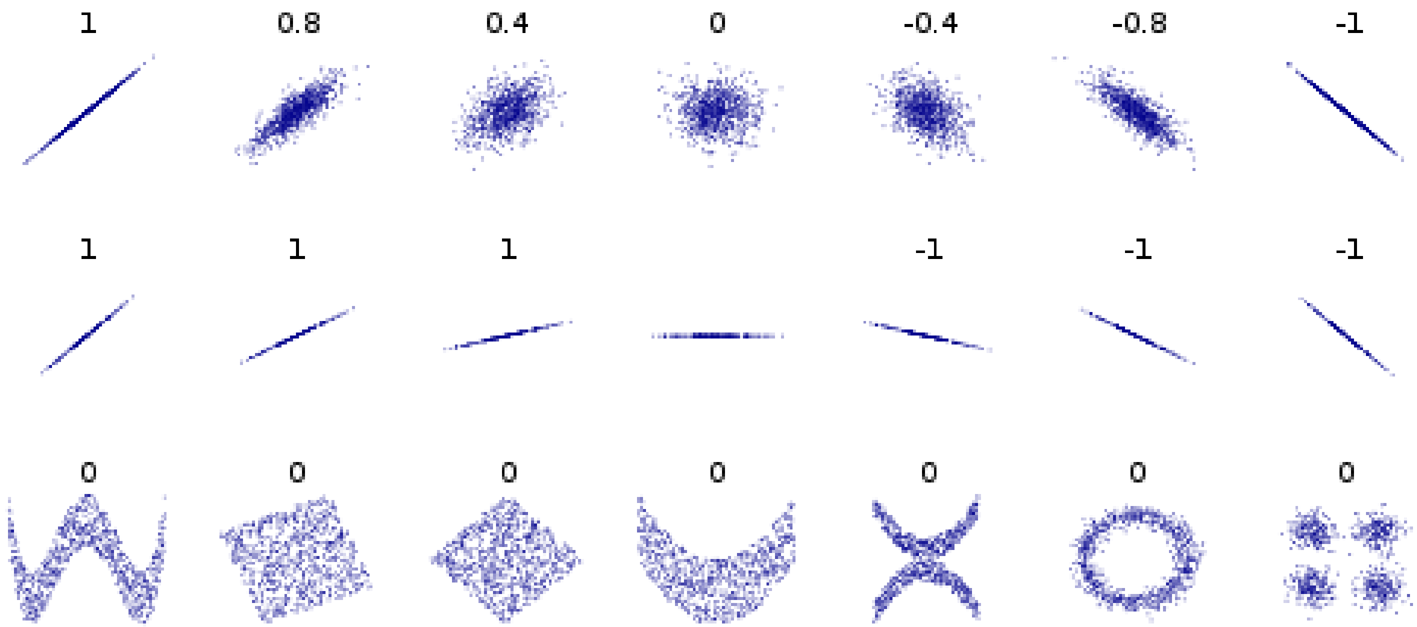
Independence and Covariance

➔Correlation:

$$\mathrm{Corr}(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

⇨Values between -1 and 1, and independent random variables have a correlation of zero

Examples



➔ **Remark:**

- ⇨ X and Y independent -> Cov(X,Y)=0;
- ⇨ X and Y dependent does not imply Cov(X,Y)=0!!! (give a look to third row)
- ⇨ Cov useful for linear relations between X and Y