Lecture 10: Random variables

In this lecture, we define random variables, the expectation, mean and standard deviation.

> A **random variable** is a function X from the probability space to the real line with the property that for every interval the set $\{X \in [a,b]\ \}$ is an event.

There is nothing complicated about random variables. They are just functions on the laboratory Ω. The reason for the difficulty in understanding random variables is solely due to the name "variable". It is not a variable we solve for. It is just a function. It quantifies properties of experiments. In any applications, the sets $X \in [a,b]$ are automatically events. The last condition in the definition is something we **do not have to worry about in general**.
If our probability space is finite, all subsets are events. In that case, **any** function on Ω is a random variable. In the case of continuous probability spaces like intervals, any piecewise continuous function is a random variable. In general, any function which can be constructed with a sequence of operations is a random variable.

1 We throw two dice and assign to each experiment the sum of the eyes when rolling two dice. For example $X[(1,2)] = 3$ or $X[(4,5)] = 9$. This random variable takes values in the set $\{2,3,4,\dots,12\}$.

> Given a random variable X, we can look at probabilities like $\mathrm{P}[\{X = 3\}\ ]$. We usually leave out the brackets and abbreviate this as $\mathrm{P}[X = 3]$. It is read as "the probability that $X = 3$."

2 Assume Ω is the set of all 10 letter sequences made of the four nucleotides G, C, A, T in a string of DNA. An example is $\omega = (G,C,A,T,T,A,G,G,C,T)$. Define $X(\omega)$ as the number of Guanin basis elements. In the particular sample ω just given, we have $X(\omega) = 3$.
Problem Assume $X(\omega)$ is the number of Guanin basis elements in a sequence. What is the probability of the event $\{X(\omega) = 2\ \}$? **Answer** Our probability space has $4^{10} = 1048576$ elements. There are 3^8 cases, where the first two elements are G. There are 3^8 elements where the first and third element is G, etc. For any pair, there are 3^8 sequences. We have $(10*9/2) = 45$ possible ways to chose a pair from the 10. There are therefore $3^8 \cdot 45$ sequences with exactly 2 amino acids G. This is the cardinality of the event $A = \{X(\omega) = 2\ \}$. The probability is $|A|/|\Omega| = 45 * 3^8/4^{10}$ which is about 0.28.

> For random variables taking finitely many values we can look at the probabilities $p_j = P[X = c_j\ ]$. This collection of numbers is called a **discrete probability distribution** of the random variable.

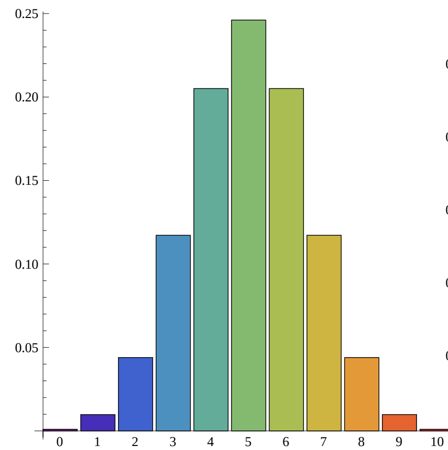
3 We throw a dice 10 times and call $X(\omega)$ the number of times that "heads" shows up. We have $\mathrm{P}[X = k\ ] = \left(\frac{10!}{k!(10-k)!}\right)/2^{10}$. because we chose k elements from $n = 10$. This distribution is called the **Binominal distribution** on the set $\{0,1,2,3,4,5,6,7,8,9,10\ \}$.

4 In the 10 nucleotid example, where X counts the number of G nucleotides, we have

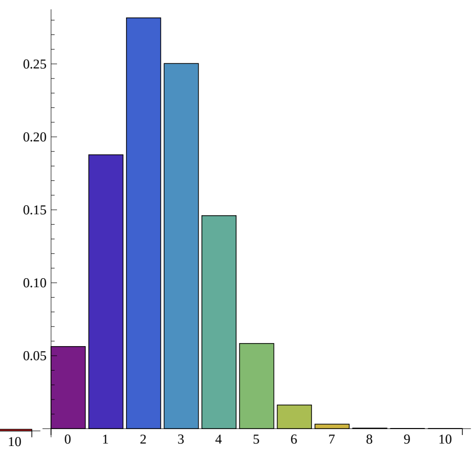
$$\mathrm{P}[X = k] = \binom{10}{k}\frac{3^{n-k}}{4^n} .$$

We can write this as $\binom{10}{k} p^k(1-p)^{n-k}$ with $p = 1/4$ and interpret it of having "heads" turn up k times if it appears with probability p and "tails" with probability $1-p$.

> If $X(k)$ counts the number of 1 in a sequence of length n and each 1 occurs with a probability p, then
> $$\mathrm{P}[X = k\ ] = \binom{n}{k} p^k(1-p)^{n-k} .$$
> This **Binomial distribution** an extremely important example of a probability distribution.



The Binominal distribution with $p = 1/2$.



The Binominal distribution with $p = 1/4$.

> For a random variable X taking finitely many values, we define the **expectation** as $m = \mathrm{E}[X] = \sum_x x\mathrm{P}[X = x\ ]$. Define the **variance** as $\mathrm{Var}[X] = \mathrm{E}[(X-m)^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$ and the **standard deviation** as $\sigma[X] = \sqrt{\mathrm{Var}[X]}$.

5 In the case of throwing a coin 10 times and head appears with probability $p = 1/2$ we have

$$\mathrm{E}[X] = 0 \cdot \mathrm{P}[X = 0] + 1 \cdot \mathrm{P}[X = 1] + 2 \cdot \mathrm{P}[X = 2] + 3 \cdot \mathrm{P}[X = 3] + \dots + 10 \cdot \mathrm{P}[X = 10] .$$

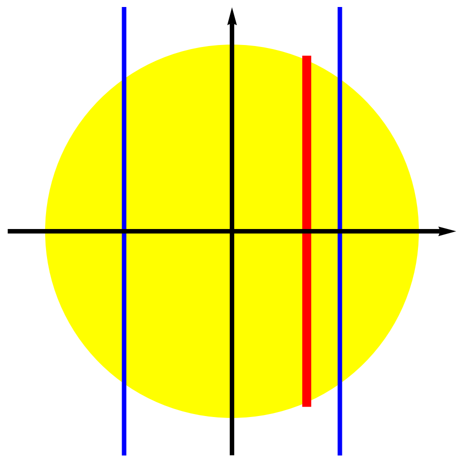
The average adds up to $10 \times p = 5$, which is what we expect. We will see next time when we discuss independence, how we can get this immediately. The variance is

$$\mathrm{Var}[X] = (0-5)^2 \cdot \mathrm{P}[X = 0\ ] + (1-5)^2 \cdot \mathrm{P}[X = 1\ ] + \dots + (10-5)^2 \cdot \mathrm{P}[X = 10\ ] .$$

It is $10p(1-p) = 10/4$. Again, we will have to wait until next lecture to see how we can get this without counting.

All these examples so far, the random variable has taken only a discrete set of values. Here is an example, where the random variable can take values in an interval. It is called a variable with a **continuous distribution**.

6 Throw a vertical line randomly into the unit disc. Let $X[\omega\ ]$ be the length of the segment cut out from the circle. What is $P[X > 1\ ]$?
Solution: we need to hit the x axes in $|x| < 1/\sqrt{3}$. Comparing lenghts gives the probability is $1/\sqrt{3}$. We have assumed here that every interval $[c,d]$ in the interval $[-1,1]$ appears with probability $(d-c)/2$.

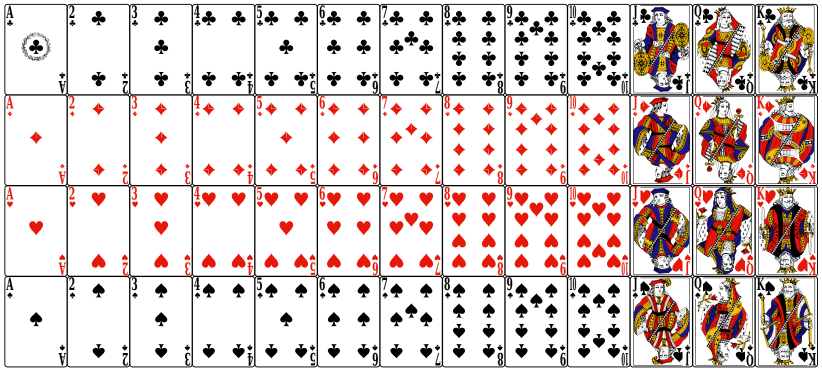


If a random variable has the property that $\mathrm{P}[X \in [a,b]] = \int_a^b f(x)\ dx$ where f is a nonnegative function satisfying $\int_{-\infty}^{\infty} f(x)\ dx = 1$. Then the expectation of X is defined as $\mathrm{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\ dx$. The function f is called the **probability density function** and we will talk about it later in the course.

In the previous example, we have seen again the Bertrand example, but because we insisted on vertical sticks, the probability density was determined. The other two cases we have seen produced different probability densities. A probability model always needs a probability function P.

Homework due February 23, 2011

1 In the card game **blackjack**, each of the 52 cards is assigned a value. You see the French card deck below in the picture. **Numbered cards 2-10** have their natural value, the **picture cards jack**, **queen**, and **king** count as 10, and aces are valued as either 1 or 11. Draw the probability distribution of the random variable X which gives the value of the card assuming that we assign to the hearts ace and diamond aces the value 1 and to the club ace and spades ace the value 11. Find the mean the variance and the standard deviation of the random variable X.



2 We look at the probability space of all 2×2 matrices, where the entries are either 1 or -1. Define the random variable $X(\omega) = \det(\omega)$, where ω is one of the matrices. The determinant is

$$\det(\begin{bmatrix} a & b \\ c & d \end{bmatrix}) = ad - bc\ .$$

Draw the probability distribution of this random variable and find the expectation as well as the Variance and standard deviation.

3 A LCD display with 100 pixels is described by a 10×10 matrix with entries 0 and 1. Assume, that each of the pixels fails independently with probability $p = 1/20$ during the year. Define the random variable $X(\omega)$ as the number of dead pixels after a year.
a) What is the probability of the event $P[X > 3]$, the probability that more than 3 pixels have died during the year?
b) What is the expected number of pixels failing during the year?