TEACHING MATHEMATICS WITH A HISTORICAL PERSPECTIVE

OLIVER KNILL

E-320: Teaching Math with a Historical Perspective **O. Knill, 2010-2022**

Lecture 8: Probability theory

8.1. Probability theory is the science of chance. It starts with **combinatorics** and leads to a theory of **stochastic processes**. Historically, probability theory initiated from gambling problems, as in **Girolamo Cardano's** gamblers manual in the 16th century. A great moment of mathematics occurred, when **Blaise Pascal** and **Pierre Fermat** jointly laid a foundation of mathematical probability theory.

8.2. It took mathematicians longer to formalize "randomness" precisely. Here is the setup as which it had been put forward by **Andrey Kolmogorov**: all possible experiments of a situation are modeled by a set Ω which is the **laboratory**. A measurable subset of experiments is called an **event**. A probability function P gives the proability $P[A]$ of an event. Measurements are done by real-valued functions X. These functions are called **random variables** and are used to **observe the laboratory**. [1]

Practice			$990.00
	min	max	current
Inside	$1	$500	$0
Outside	$1	$500	$27

FIGURE 1. Probability theory started with gambling (Cardano).

[1] If Ω is finite, then every subset of Ω can be an event and every function X can be random variable. In general, there can be subsets of Ω which can not be assigned a probability in a reasonable way.

8.3. As an example, let us model the process of throwing a coin 5 times. An **experiment** is a word like *ttht*, where h stands for “head” and t represents “tail”. The laboratory consists of all possible 32 words. We could look for example look the event A where the first two coin tosses are tail. It is the set $A = \{ttttt, tttth, tttht, ttthh, tthtt, tthth, tthht, tthhh\}$. The most natural probability function is $P[A] = |A|/|\Omega|$, in this case $P[A] = 8/32 = 1/4$. We could also look at the random variable X which assigns to a word w the number of heads in w. For every experiment, we get a value, like for example, $X[tthht] = 2$.

8.4. In order to make precise statements about randomness, the specification of the **probability measure** is important. This is a function P from the set of all events to the interval $[0, 1]$. It should have the property that $P[\Omega] = 1$ and $P[A_1 \cup A_2 \cup \dots] = P[A_1] + P[A_2] + \cdots$, if A_i are disjoint events.

8.5. The most natural probability measure on a finite set Ω is $P[A] = \|A\|/\|\Omega\|$, where $\|A\|$ stands for the number of elements in A. It is the “”number of good cases” divided by the “number of all cases”. For example, to count the probability of the event A that we throw 3 heads during the 5 coin tosses, we have $|A| = 10$ possibilities. Since the entire laboratory has $|\Omega| = 32$ possibilities, the probability of the event is $P[A] = 10/32$. In order to study these probabilities, **combinatorics** helps:

How many ways are there to:	**The answer is:**
rearrange or permute n elements	$n! = n(n-1)...2 \cdot 1$
choose k from n with repetitions	n^k
pick k from n if order matters	$\frac{n!}{(n-k)!}$
pick k from n with order irrelevant	$\left(\begin{array}{c} n \\ k \end{array}\right) = \frac{n!}{k!(n-k)!}$

8.6. The **expectation** $\mathrm{E}[X]$ of a random variable X is defined as the sum $m = \sum_{\omega \in \Omega} X(\omega)P[\{\omega\}]$. In our coin toss experiment, this is $5/2$. The **variance** of X is the expectation of $(X - m)^2$. In our coin experiments, it is $5/4$. Its square root is called the **standard deviation**. This is the expected deviation from the mean. An event happens **almost surely** if the event has probability 1.

8.7. An important case of a random variable is $X(\omega) = \omega$ on $\Omega = R$ equipped with probability $P[A] = \int_A \frac{1}{\sqrt{\pi}} e^{-x^2}\ dx$, the **standard normal distribution**. Analyzed first by **Abraham de Moivre** in 1733, it was studied by **Carl Friedrich Gauss** in 1807 and therefore also called **Gaussian distribution**.

8.8. Two random variables X, Y are called **decorrelated**, if $E[XY] = E[X] \cdot E[Y]$. If for any functions f, g also $f(X)$ and $g(Y)$ are decorrelated, then X, Y are called **independent**. Two random variables are said to have the same distribution, if for any $a < b$, the events $\{a \leq X \leq b\ \}$ and $\{a \leq Y \leq b\ \}$ are independent. If X, Y are decorrelated, then the relation $\mathrm{Var}[X] + \mathrm{Var}[Y] = \mathrm{Var}[X + Y]$ holds. This is just the **Pythagorean theorem**, because decorrelated can be understood geometrically: $X - E[X]$ and $Y - E[Y]$ are orthogonal.

8.9. A common problem is to study the sum of independent random variables X_n with identical distribution. One abbreviates this with **IID**. Here are the three important theorems which we formulate in the case, whee all random variables are assumed to have expectation 0 and standard deviation 1. Let $S_n = X_1 + \cdots + X_n$ be the n'th sum of the IID random variables. It is also called a **random walk**.

LLN **Law of Large Numbers** assures that S_n/n converges to 0.

CLT **Central Limit Theorem**:$S_n/\sqrt{n}$ approaches the Gaussian distribution.

LIL **Law of Iterated Logarithm:** $S_n/\sqrt{2n \log\log(n)}$ accumulates in $[-1, 1]$.

(For LLN and LIL, one should say that the converges happens with probability 1. For the CLT, the convergence is in the sense of distributions.)

8.10. The LLN shows that one can find out about the expectation by averaging experiments.
The CLT explains why one sees the standard normal distribution so often.
The LIL gives us a precise estimate how fast S_n grows.
Things become interesting if the random variables are no more independent. Generalizing LLN,CLT,LIL to such situations is part of ongoing research.

8.11. Here is an open questions in probability theory:

Are $\pi, e, \sqrt{2} \cdots$ normal in the following sense: do all digits appear with the same frequency?

8.12. **Statistics** is the science of modeling random events in a probabilistic setup. Given data points, we want to find a **model** which fits the data best. This allows to **understand the past**, **predict the future** or **discover laws of nature**. The most common task is to find the **mean** and the **standard deviation** of some data. The mean is also called the **average** and given by $m = \frac{1}{n} \sum_{k=1}^{n} x_k$. The variance is $\sigma^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - m)^2$ with standard deviation σ.

8.13. A sequence of random variables X_n produce a **stochastic process**. Continuous versions of such processes are where X_t is a curve of random random variables. An important example is **Brownian motion**, which is a model of a random particles.

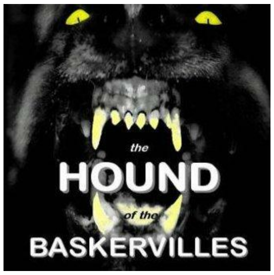
8.14. Besides gambling and analyzing data, also **physics** has been an important motor to develop probability theory. An example is statistical mechanics where laws of nature are studied with probabilistic methods. A famous physical law is **Ludwig Boltzmann's** relation $S = k \log(W)$ for entropy, a formula which decorates Boltzmann's tombstone. The **entropy** of a probability measure $P[\{k\}] = p_k$ on a finite set $\{1, ..., n\}$ is defined as $S = -\sum_{i=1}^{n} p_i \log(p_i)$. Today, we would reformulate Boltzmann's law and say that it is the expectation $S = E[\log(W)]$ of the logarithm of the "Wahrscheinlichkeit" random variable $W(i) = 1/p_i$ on $\Omega = \{1, ..., n\ \}$.

8.15. Entropy is important because nature tries to maximize it In the simplest situations, if a system has n states, the $W = n$. If each state has an energy E_i and the free energy $S - E$ is minimized, the probability distribution $p_i = e^{-E_i}/Z$ is the **Boltzman distribution**.

Here are the most important combinatorics problems:

How many ways are there to:	**The answer is:**
permute n elements	$n! = n(n-1)...2 \cdot 1$
choose k from n with repetitions	n^k
pick k different from n if order matters	$\frac{n!}{(n-k)!}$
pick k different from n where order does not matter	$\begin{pmatrix} n \\ k \end{pmatrix} = \frac{n!}{k!(n-k)!}$

Problem 1: You play ”scrabble”. You are stuck with the letters $STORY$. How many single words of length 5 can you write?



Problem 2: the "Hound of Baskervilles" has 338'787 letters. How many novels are there with this number of letters? You can assume an alphabet of 30 including space and punctuations.

Problem 3: How many ways are there to chose 3 people from a contestant group of 12 if the order does not matter?

Problem 4: A combination lock has 40 numbers $0-39$. A lock combination consists of 3 different numbers, where the order matters. How many different lock combinations are there?

Work problems

8.16. 1) We want to understand the famous **Monty Hall problem**

You have to choose from three doors. Behind one door is a car and behind the others are goats. You pick a door. The host, who knows what's behind the doors, opens an another door, one which has a goat. You have the choice to choose the door or to switch. What is better?

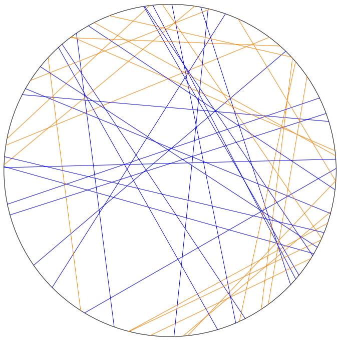
The problem became sensation and controversy in 1991. Intuitive argumentation can lead to the conclusion that it does not matter whether to change the door or not. When asked, a large

majority of test persons tell that it does not matter. a) We first assume that we decide not to switch.
You choose a door. Note that the revelation of the host does not affect your choice.
What is the probability that you win in this case?
b) Now we switch. We look at three possibilities now.
What happens if you initially chose the door with the car? Do you win or lose in this case?
c) What happens if you initially chose the door with the goat? Do you win or lose in this case?
d) What do you conclude

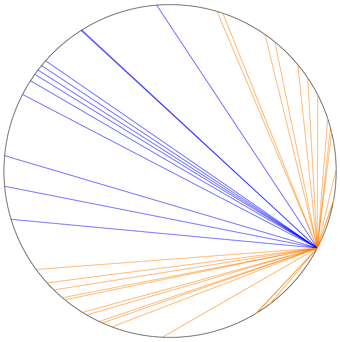
8.17. 2) For the following question, most people would say 1/2.

Dave has 2 kids. One of them is a boy. What is the probability that the other kid is a girl?

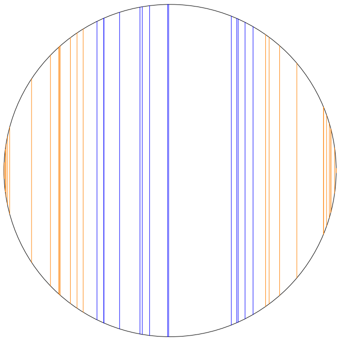
8.18. The **Bertrand paradox** illustrates that one has to be clear on how to setup a probabilistic model in a concrete situation.



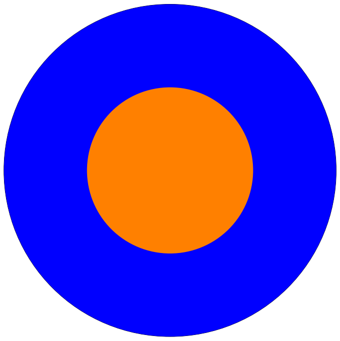
We throw random lines onto the unit disc. What is the probability that the line intersects the disc with a length $\geq \sqrt{3}$, the length of the inscribed equilateral triangle?



Problem 1) Take an arbitrary point on the boundary of the disc. The set of all lines through that point are parameterized by an angle ϕ. For which midpoints is the length of the chord longer than the equilateral triangle length? By comparing angles, what is the probability?



Problem 2) Take now all lines perpendicular to a fixed diameter. The entire diameter has length 2. Where do the chords hit the diameter so that it is longer than $\sqrt{3}$? By comparing lengths, what is the probability?



Problem 3) Look at the midpoints of the chords. Where does such a midpoint have to be so that the chord is longer than $\sqrt{3}$? By comparing areas, what is the probability now?

8.19. The origins of probability was in gambling. We look here closely at the Petersburg paradox, which had been devised by Daniel Bernoulli in 1738. You pay a fixed entrance fee C and you get the prize 2^T, where T is the number of times, the casino flips a coin until "head" appears.
For example, you enter 10 dollars. If the sequence of coin experiments would give "tail, tail, tail, head", you win $2^3 - 10 = 8 - 10 = -2$ dollars. This means you have lost 2 dollars in this game.

a) Build groups of 2-4. One is the casino, the others play the casino. Choose an entrance fee which you think is fair and play as many times as time allows. In the end, record your winning.
b) What is the probability that you lose your entire winning? That is, what is the chance that we have "head" the first time? Note that $T = 0$ in this case. c) What is the probability that we have "head" the second time? Note that $T = 2$ in this case. How much do we win or lose in this case?
d) What is the probability that "head" appears the third time? Note that $T = 3$ in this case. How much did you win or lose in this case?
e) What is the probability that "head" appears at time $T = n$ the first time? How much did you win or lose in this case?
Fair would be an entrance fee which is equal to the expectation of the win, which is $1 \cdot P[T = 0] + 2 \cdot P[T = 1] + 5 \cdot P[T = 2] + ...$ What does "fair" mean? For example, the situation $T = 20$ is so improbable that it never occurs in the life-time of a person. Therefore, for any practical reason, one has not to worry about large values of T. This, as well as the finiteness of money resources is the reason, why casinos do not have to worry about the following bullet proof **martingale strategy** in roulette: bet c dollars on red. If you win, stop, if you lose, bet $2c$ dollars on red. If you win, stop. If you lose, bet $4c$ dollars on red. Keep doubling the bet. Eventually after n steps, red will occur and you will win $2^n c - (c + 2c + \cdots + 2^{n-1}c) = c$ dollars.

8.20. Here is some additional information. How does one solve the Petersburg paradox? What would be a reasonable entrance fee in "real life"? Bernoulli proposed to replace the expectation $\mathrm{E}[G]$ of the profit $G = 2^T$ with the expectation $(\mathrm{E}[\sqrt{G}])^2$, where $u(x) = \sqrt{x}$ is called a **utility function**. This would lead to a fair entrance

$$(\mathrm{E}[\sqrt{G}])^2 = (\sum_{k=1}^{\infty} 2^{k/2} 2^{-k})^2 = \frac{1}{(\sqrt{2}-1)^2} \sim 5.828... \, .$$

Similar effects appear in political situations as in **voting systems**, where different voting systems can produce different winners. The following example is by Donald Saari:

"Consider 15 people deciding what beverage to serve at a party. Six prefer milk first, wine second, and beer third; five prefer beer first, wine second, and milk third; and four prefer wine first, beer second, and milk third. In a plurality vote, milk is the clear winner. But if the group decides instead to hold a runoff election between the two top contenders milk and beer, then beer wins, since nine people prefer it over milk. And if the group awards two points to a drink each time a voter ranks it first and one point each time a voter ranks it second, suddenly wine is the winner."