Chapter 3. PROBABILITY THEORY BASICS I

Probability theory took almost 300 years to develop from the ideas of Pascal, Fermat and Huygens, around 1650, to the currently acceptable axiomatic definition of Kolmogorov.

Intuitively, probability of an event E is the ratio between the number of favorable elementary events involved in E to the sum of all possible elementary events involved in E.

Any probabilistic statement must refer to an underlying probability space.

A **probability space** is defined in terms of a **sample space** Ω (often with an algebraic structure – for example, outcomes of cube tossing) and a **probability measure** (*probability distribution*) defined on Ω.

Subsets of a sample space Ω are called **events**. Elements of Ω are referred to as **elementary events**.

Intuitively, the sample space represents the set of all possible outcomes of a **probabilistic experiment** – for example of a cube tossing. An event represents a collection of possible outcomes.

Intuitively, probability of an event E is the ration between the number of (favorable) elementary events involved in E to the number of all possible elementary evens.

Axiomatic approach Probability distribution on a set Ω is every function $\Pr : 2^{\Omega} \rightarrow [0,1]$, satisfying axioms:

1. $\Pr(\{x\}) \geq 0$ for any elementary event x;
2. $\Pr(\Omega) = 1$
3. $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ if $A, B \subseteq S$, $A \cap B = \emptyset$.

Example: Probabilistic experimentr – cube tossing; elementary events – outcomes of cube tossing; probability distribution – $p_1, p_2, p_3, p_4, p_5, p_6$, $\sum_{i=1}^{6} p_i = 1$.

In general, a sample space is an arbitrary set. However, often we need (wish) to consider only some (family) of all possible events of 2^{Ω}.

The fact that not all collections of events lead to well-defined probability space leads to the concepts presented on the next slide.

PROBABILITY SPACES

Definition 1 *A* ***σ-field*** $(\Omega, \mathbf{F})$ *consists of a sample space* Ω *and a collection* $\mathbf{F}$ *of subsets of* Ω *satisfying the following conditions:*

1. $\emptyset \in \mathbf{F}$
2. $\varepsilon \in \mathbf{F} \Rightarrow \overline{\varepsilon} \in \mathbf{F}$
3. $\varepsilon_1, \varepsilon_2, \ldots \in \mathbf{F} \Rightarrow (\varepsilon_1 \cup \varepsilon_2 \cup \ldots) \in \mathbf{F}$

Consequence 2 *A* σ-field *is closed under countable unions and intersections.*

Definition 3 *A* **probability measure** (distribution) $Pr : \mathbf{F} \rightarrow \mathbf{R}^{\geq 0}$ *on a* σ*-field* $(\Omega,\mathbf{F})$ *is a function satisfying conditions:*

1. *If* $\varepsilon \in \mathbf{F}$, *then* $0 \leq Pr(\varepsilon) \leq 1$.
2. $Pr[\Omega] = 1$.
3. *For mutually disjoint events* $\varepsilon_1, \varepsilon_2, \ldots$
   $Pr[\bigcup_i \varepsilon_i] = \sum_i Pr(\varepsilon_i)$

Definition 4 *A* **probability space** $(\Omega,\mathbf{F},Pr)$ *consists of a* σ*-field* $(\Omega,\mathbf{F})$ *with a probability measure* Pr *defined on* $(\Omega, \mathbf{F})$.

PROBABILITIES and their PROPERTIES

Properties 5

$$\begin{aligned} Pr(\overline{\varepsilon}) &= 1 - Pr(\varepsilon) \\ Pr(\varepsilon_1 \cup \varepsilon_2) &= Pr(\varepsilon_1) + Pr(\varepsilon_2) - Pr(\varepsilon_1 \cap \varepsilon_2) \\ Pr(\bigcup_{i \geq 1} \varepsilon_i) &\leq \sum_{i \geq 1} Pr(\varepsilon_i). \end{aligned}$$

Definition 6 **Conditional probability** *of an event* ε_1 *given an event* ε_2 *is defined by*

$$Pr[\varepsilon_1 | \varepsilon_2] = \frac{Pr[\varepsilon_1 \cap \varepsilon_2]}{Pr[\varepsilon_2]}$$

if $Pr[\varepsilon_2] > 0$.

Theorem 7 *Law of total probability Let* $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k$ *be a partition of the sample space* Ω. *Then for any event* ε

$$Pr[\varepsilon] = \sum_{i=1}^{k} Pr[\varepsilon | \varepsilon_i] \cdot Pr[\varepsilon_i]$$

Theorem 8 ***(Bayes' Rule/Law)***

(a) $Pr(\varepsilon_1) \cdot Pr(\varepsilon_2|\varepsilon_1) = Pr(\varepsilon_2) \cdot Pr(\varepsilon_1|\varepsilon_2)$ *basic equality*

(b) $Pr(\varepsilon_2|\varepsilon_1) = \frac{Pr(\varepsilon_2)\, Pr(\varepsilon_1|\varepsilon_2)}{Pr(\varepsilon_1)}$ *simple version*

(c) $Pr[\varepsilon_1|\varepsilon] = \frac{Pr[\varepsilon_1 \cap \varepsilon]}{Pr[\varepsilon]} = \frac{Pr[\varepsilon|\varepsilon_1] \cdot Pr[\varepsilon_1]}{\sum_{i=1}^{k} Pr[\varepsilon|\varepsilon_i] \cdot Pr[\varepsilon_i]}.$ *extended version*

Definition 9 Independence

1. *Two events* $\varepsilon_1, \varepsilon_2$ *are called* **independent** *if*

$$Pr(\varepsilon_1 \cap \varepsilon_2) = Pr(\varepsilon_1) \cdot Pr(\varepsilon_2)$$

.

2. *A collection of events* $\{\varepsilon_i | i \in I\}$ *is* **independent** *if for all subsets* $S \subseteq I$

$$Pr\left[\bigcap_{i \in S} \varepsilon_i\right] = \prod_{i \in S} Pr[\varepsilon_i].$$

MODERN (BAYESIAN) INTERPRETATION of BAYES RULE

for the entire process of learning from evidence has the form

$$Pr[\varepsilon_1|\varepsilon] = \frac{Pr[\varepsilon_1 \cap \varepsilon]}{Pr[\varepsilon]} = \frac{Pr[\varepsilon|\varepsilon_1] \cdot Pr[\varepsilon_1]}{\sum_{i=1}^{k} Pr[\varepsilon|\varepsilon_i] \cdot Pr[\varepsilon_i]}.$$

In modern terms the last equation says that $Pr[\varepsilon_1|\varepsilon]$, the probability of a hypothesis ε_1 (given information ε), equals $\Pr(\varepsilon_1)$, our initial estimate of its probability, times $Pr[\varepsilon|\varepsilon_1]$, the probability of each new piece of information (under the hypothesis ε_1), divided by the sum of the probabilities of data in all possible hypotheis (ε_i).

TWO BASIC INTERPRETATIONS of PROBABILITY

In Frequentist interpretation , probability is defined with respect to a large number of trials, each producing one outcome from a set of possible outcomes - the probability of an event A **, Pr(A), is a proportion of trials producing an outcome in A.**

In Bayesian interpretation probbability measures a degree of belief. Bayes' theorem then links the degree of belief in a proposition before and after receiving a dditional evidence that the proposition holds.

EXAMPLE 1

Let us toss a two regular cubes, one after another and let

$$\varepsilon_1 \ \textbf{ be the event that the sum of both tosses is } \geq 10$$
$$\varepsilon_2 \ \textbf{ be the event that the first toss providdes } 5$$

Then, how much are: $\boldsymbol{Pr}(\varepsilon_1), \boldsymbol{Pr}(\varepsilon_2), \boldsymbol{Pr}(\varepsilon_1|\varepsilon_2), \boldsymbol{Pr}(\varepsilon_1 \cap \varepsilon_2)$**?**

$$\boldsymbol{Pr}(\varepsilon_1) = \frac{6}{36}$$
$$\boldsymbol{Pr}(\varepsilon_2) = \frac{1}{6}$$
$$\boldsymbol{Pr}(\varepsilon_1|\varepsilon_2) = \frac{2}{6}$$
$$\boldsymbol{Pr}(\varepsilon_1 \cap \varepsilon_2) = \frac{2}{36}$$

EXAMPLE 2

Three coins are given - two fair ones and in the third one heads land with probability 2/3, but we do not know which one is not fair one.

When making an experiment and flipping all coins let the first two come up heads and the third one comes up tails. What is probability that the first coin is the biased one?

Let ε_i be the event that the ith coin is biased and B be the event that three coins flips came up heads, heads, tails.

Before flipping coins we have $Pr(\varepsilon_i) = \frac{1}{3}$ for all i. After flipping coins we have

$$Pr(B|\varepsilon_1) = Pr(B|\varepsilon_2) = \frac{2}{3}\frac{1}{2}\frac{1}{2} = \frac{1}{6}$$

$$Pr(B|\varepsilon_3) = \frac{1}{2}\frac{1}{2}\frac{1}{3} = \frac{1}{12}$$

and using Bayes' law we have

$$Pr(\varepsilon_1|B) = \frac{Pr(B|\varepsilon_1)Pr(\varepsilon_1)}{\sum_{i=1}^{3} Pr(B|\varepsilon_i)Pr\,(\varepsilon_i)} = \frac{\frac{1}{6}\cdot\frac{1}{3}}{\frac{1}{6}\cdot\frac{1}{3}+\frac{1}{6}\cdot\frac{1}{3}+\frac{1}{12}\cdot\frac{1}{3}} = \frac{2}{5}$$

Therefore, the outcome of the three coin flips increases the likehood that the first coin is biased from 1/3 to 2/5

THEOREM

Let A and B be two events and let $\Pr(B) \neq 0$. Events A and B are independent if and only if

$$\Pr(A|B) = \Pr(A).$$

Proof

- Assume that A and B are independent and $\Pr(B) \neq 0$. By definition we have

  $$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

  and therefore

  $$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(A) \cdot \Pr(B)}{\Pr(B)} = \Pr(A).$$

- Assume that $\Pr(A|B) = \Pr(A)$ and $\Pr(B) \neq 0$. Then

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

  and multiplying by $\Pr(B)$ we get

  $$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

  and so A and B are independent.

SUMMARY

- The notion of conditional probability, of A given B, was introduced in order to get an instrument for analyzing an experiment A when one has partial information B about the outcome of the experiment A before experiment has finished.
- **We say that two events A and B are independent if the probability of A is equal to the probability of A given B,**
- Other fundamental instruments for analysis of probabilistic experiments are **random variables as functions from the sample space to R**, and the expectation of a random variable as the weighted average of the values of the random variable.

MONTY HALL PARADOX

Let us assume that you see three doors D1, D2 and D3 and you know that behind one door is a car and behind other two are goats.

You have a chance to choose one door. If it is door with car behind the car is yours, if it is door with a goat behind you will have to milk the goat for one year.

Which door you will choose?

Let us now assume that you have chosen the door D1.

Now comes a moderator who knows where car is and opens one of the doors D_2 or D_3, say D2, and you see that the goat is in.

Let us assume that you now get a chance to change your choise of the door.
Should you do that?

Let C_1 denote the event that the car is behind the door D1.
Let C_3 denote the event that the car is behind the door D3.
Let M_2 denote the event that moderator openes the door D2.

Let us assume that the moderator choosed a door at random if goats were behind both doors he could open. In such a case we have

$$Pr[C_1] = \frac{1}{3} = Pr[C_3] \quad Pr[M_2|C1] = \frac{1}{2}, \qquad Pr[M_2|C_3] = 1$$

Then it holds

$$Pr[C_1|M_2] = \frac{Pr[M_2|C_1]Pr[C_1]}{Pr[M_2]} = \frac{Pr[M_2|C_1]Pr[C_1]}{Pr[M_2|C_1]Pr[C_1] + Pr[M_2|C_3]Pr[C_3]} = \frac{1/6}{1/6 + 1/3} = \frac{1}{3}$$

Similarly

$$Pr[C_3|M_2] = \frac{Pr[M_2|C_3]Pr[C_3]}{Pr[M_2]} = \frac{Pr[M_2|C_3]Pr[C_3]}{Pr[M_2|C_1]Pr[C_1] + Pr[M_2|C_3]Pr[C_3]} = \frac{1/3}{1/6 + 1/3} = \frac{2}{3}$$

DISTRIBUTION FUNCTIONS

Definition 10 ***A*** random variable X ***is a real valued function over the sample space*** Ω ***[of a*** σ***-field*** $(\Omega, \mathbf{F})$***], that is*** $X : \Omega \rightarrow \mathbf{R}$***, such that for all*** $x \in \mathbf{R}$***:***

$$\{\omega \in \Omega | X(\omega) \leq x\} \in \mathbf{F}.$$

Definition 11 ***The*** distribution function $F_X : \mathbf{R} \rightarrow [0, 1]$ ***of a random variable*** X ***is defined as***

$$F_X(x) = Pr[X \leq x].$$

The density function $p : \mathbf{R} \rightarrow [0, 1]$ ***for a random variable*** X ***is defined as***

$$p_X(x) = Pr[X = x].$$

Example 12 ***Let Ω_0 be the set of all possible outcomes of throwing simultaneously three dices.*** $|\Omega_0| =???$
Let $X(\omega)$ $(\omega \in \Omega_0)$ be the sum of numbers on dices. For the density function $p_X(x)$ it holds

X	**3**	**4**	**5**	**6**	**7**	**8**	**9**	...
$p_X(x)$	$\frac{1}{216}$	$\frac{3}{216}$	$\frac{6}{216}$	$\frac{10}{216}$	$\frac{12}{216}$	$\frac{21}{216}$	$\frac{25}{216}$	...

Definition 13 ***The*** joint distribution function $F_{X,Y} : \mathbf{R} \times \mathbf{R} \rightarrow [0, 1]$ ***for random variables X and Y is defined by***

$$F_{X,Y}(x, y) = Pr[\{X \leq x\} \wedge \{Y \leq y\}]$$

The joint density function $Pr_{X,Y} : \mathbf{R} \times \mathbf{R} \rightarrow [0, 1]$ ***for random variables X and Y is defined by***

$$Pr_{X,Y}(x, y) = Pr[\{X = x\} \wedge \{Y = y\}]$$

Definition 14 ***Two random variables X, Y are called*** independent ***if***

$$x, y \in \mathbf{R} \Rightarrow Pr_{X,Y}(x, y) = Pr[X = x] \cdot Pr[Y = y]$$

RANDOM VARIABLES - A SIMPLER APPROACH

A random variable is a function defined on the elementary events of a probability space.

Example In case of two tosses of a dice, the value of a random variable V can be the sum of the two spots on the dice rolls.

The value of S can therefore be an integer from the interval $[2, 12]$.

A random variable V with n potential values $v_1, v_2, \ldots, v_n$ is characterized by a probability distribution $p = (p_1, p_2, \ldots, p_n)$, where p_i is probability that V takes the value v_i.

The concept of random variable is one of the most important of modern science and technology. '

EXPECTATION – MEAN of RANDOM VARIABLES

Definition 15 *The* **expectation** (**mean** or **expected value**) $\mathbf{E}[X]$ *of a random variable* X *is defined as*

$$\mathbf{E}[X] = \sum_{\omega \in \Omega} X(\omega) Pr_X(\omega).$$

Properties 16

$$\begin{aligned} \mathbf{E}[X+Y] &= \mathbf{E}[X] + \mathbf{E}[Y]. \\ \mathbf{E}[c \cdot X] &= c \cdot \mathbf{E}[X]. \\ \mathbf{E}[X \cdot Y] &= \mathbf{E}[X] \cdot \mathbf{E}[Y], \qquad \text{if } X, Y \text{ are independent} \end{aligned}$$

The first of the above equalities is known as **linearity of expectations**. It can be extended to a finite number of random variables $X_1, \ldots, X_n$ to hold

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

and also to any countable set of random variables $X_1, X_2, \ldots$ to hold:

If $\sum_{i=1}^{\infty} \mathbf{E}[|X_i|] < \infty$, then $\sum_{i=1}^{\infty} |X_i| < \infty$ and

$$\mathbf{E}[\sum_{i=1}^{\infty} X_i] = \sum_{i=1}^{\infty} \mathbf{E}[X_i].$$

EXPECTATION VALUES

For any random variable X let R_X be the set of values of X. Using R_X one can show that

$$E[X] = \sum_{x \in \mathbf{R}_X} x \cdot Pr(X = x).$$

Using that one can show that for any real a, b it holds

$$\begin{aligned} \mathbf{E}[aX + b] &= \sum_{x \in \mathbf{R}_X} (ax + b)\mathrm{Pr}(X = x) \\ &= a \sum_{x \in \mathbf{R}_X} x \cdot \mathrm{Pr}(X = x) = b \sum_{x \in \mathbf{R}_X} \mathrm{Pr}(X = x) \\ &= a \cdot \mathbf{E}[X] + b \end{aligned}$$

The above relation is called weak linearity of expectation.

JENSEN's INEQUALITY

Example **We show taht for any random variable X it holds**

$$\mathbf{E}[X^2] \geq (\mathbf{E}[X])^2$$

Indeed, let $Y = (X - \mathbf{E}[X])^2$

$$\begin{aligned} 0 \leq \mathbf{E}[Y] &= \mathbf{E}[(X - \mathbf{E}[X])^2] \\ &= \mathbf{E}[X^2 - 2X\mathbf{E}[X] + (\mathbf{E}[X])^2] \\ &= \mathbf{E}[X^2] - 2\mathbf{E}[X\mathbf{E}[X]] + (\mathbf{E}[X])^2 \\ &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2 \end{aligned}$$

This can be generalised as follows

Theorem - Jensen's inequality **If f is a convex function, then**

$$\mathbf{E}[f(X)] \geq f(\mathbf{E}[X])$$

INDICATOR VARIABLES

A random variable X is called an indicator variable if X takes on only values 1 and 0.

For any set $A \subset S$, one can define an indicator variable X_A that takes value 1 on A and 0 on $S - A$, if $(S, \mathbf{Pr})$ is the underlying probability space.

It holds

$$
\begin{aligned}
\mathbf{E}[X_A] &= \sum_{s \in S} X_A(s) \cdot \Pr(\{s\}) \\
&= \sum_{s \in A} X_A(s) \cdot \Pr(\{s\}) + \sum_{s \in S-A} X_A(s) \cdot \Pr(\{s\}) \\
&= \sum_{s \in A} 1 \cdot \Pr(\{s\}) + \sum_{s \in S-A} 0 \cdot \Pr(\{s\}) \\
&= \sum_{s \in A} \Pr(\{s\}) \\
&= \Pr(A)
\end{aligned}
$$

VARIANCE and STANDARD DEVIATION

Definition 17 *For a random variable X* **variance** $\mathbf{V}X$ *and* **standard deviation** σX *are defined by*

$$\mathbf{V}X = \mathbf{E}((X - \mathbf{E}X)^2)$$
$$\sigma X = \sqrt{\mathbf{V}X}$$

Since

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

it holds

$$\mathbf{V}X = \mathbf{E}(X^2) - (\mathbf{E}X)^2$$

Example 18 $\Omega = \{1, 2, \ldots, 10\}$, $Pr(i) = \frac{1}{10}$, $X(i) = i$; $Y(i) = i - 1$ *if* $i \leq 5$ *and* $Y(i) = i + 1$ *otherwise.*

$\mathbf{E}X = \mathbf{E}Y = 5.5$, $\mathbf{E}(X^2) = \frac{1}{10}\sum_{i=1}^{10} i^2 = 38.5$, $\mathbf{E}(Y^2) = 44.5$

$\mathbf{V}X = 8.25$, $\mathbf{V}Y = 14.25$

For independent random variables X and Y and a real number c it holds

- $\mathbf{V}(cX) = c^2\mathbf{V}(X)$; $\sigma(cX) = c\sigma(X)$
- $\mathbf{V}(X + Y) = \mathbf{V}(X) + \mathbf{V}(Y)$. $\sigma(X + Y) = \sqrt{V(X) + V(Y)}$.

MOMENTS

Definition 19 ***For*** $k \in \mathbf{N}$ ***the k-th moment*** m_X^k ***and the k-th central moment*** μ_X^k ***of a random variable*** X ***are defined as follows***

$$m_X^k = \mathbf{E}X^k$$
$$\mu_X^k = \mathbf{E}((X - \mathbf{E}X)^k)$$

For a random variable X **its mean is sometimes denoted by** $\mu_X = m_X^1$ **and its variance is actually** μ_X^2**.**

EXAMPLE I

Each week there is a lottery that always sells 100 tickets. One ticket wins 100 milions, oher tickets win nothing.

What is better: to buy in one week two tickets (Strategy I) or two tickets in two different weeks (Strategy II)?

Or none of these two strategies is better than the second one?

EXAMPLE II

At Strategy I we win (in millions)

$$0 \textbf{ with probability } 0.98$$

$$100 \textbf{ with probability } 0.02$$

At Strategy II we win (in millions)

$$0 \textbf{ with probability } 0.9801 = 0.99 \cdot 0.99$$

$$100 \textbf{ with probability } 0.0198 = 2 \cdot 0.01 \cdot 0.99$$

$$200 \textbf{ with probability } 0.0001 = 0.01 \cdot 0.01$$

Variance at Strategy I is 196

Variance at Strategy II is 198

PROBABILITY GENERATING FUNCTION

The probability density function of a random variable X whose values are natural numbers can be represented by the following *probability generating function* (PGF):

$$G_X(z) = \sum_{k\geq 0} Pr(X=k)\cdot z^k.$$

Properties 20

$$G_X(1) = 1$$

$$\mathbf{E}X = \sum_{k\geq 0} k\cdot Pr(X=k) = \sum_{k\geq 0} Pr(X=k)\cdot (k\cdot 1^{k-1}) = G'_X(1).$$

Since it holds

$$\begin{aligned}\mathbf{E}(X^2) &= \sum_{k\geq 0} k^2\cdot Pr(X=k)\\ &= \sum_{k\geq 0} Pr(X=k)\cdot (k\cdot (k-1)\cdot 1^{k-2} + k\cdot 1^{k-1})\\ &= G''_X(1) + G'_X(1)\end{aligned}$$

we have

$$\mathbf{V}X = G''_X(1) + G'_X(1) - (G'_X(1))^2.$$

WHY ARE PGF USEFUL?

For many important probability distributions their PGF are very simple and easy to work with.

For a uniform distribution on the set $\{0, 1, \ldots, n-1\}$ the PGF has form

$$U_n(z) = \frac{1}{n}(1 + z + \ldots + z^{n-1}) = \frac{1}{n} \cdot \frac{1 - z^n}{1 - z}.$$

Problem is with the case $z = 1$.

WAY OUT

If

$$G(z)=\sum_{n\geq 0} g_n z^n$$

is any power series that converges for at least one value of z **with** $|z|>1$**, then this property have also** $G^{'}(z)=\sum_{n\geq 0} n g_i z^{n-1}$ **and** $G^{''}(z)$**,** $G^{'''}(z)$**. By Taylor theorem we have**

$$G(1+t)=G(1)+\frac{G^{'}(1)}{1!}t+\frac{G^{''}(1)}{2!}t^2+\ldots.$$

Therefore

$$U_n(1+t)=\frac{1}{n}\frac{(1+t)^n-1}{t}=\frac{1}{n}\binom{n}{1}+\frac{1}{n}\binom{n}{2}t+\frac{1}{n}\binom{n}{3}t^2+\ldots+\frac{1}{n}\binom{n}{n}t^{n-1}.$$

and consequently

$$U_n(1)=1,\quad U_n^{'}(1)=\frac{n-1}{2}\quad U_n^{''}(1)=\frac{(n-1)(n-2)}{2}.$$

PROPERTIES of GENERATING FUNCTIONS

Properties 21 *If $X_1, \ldots, X_k$ are independent random variables with PGFs $G_1(z), \ldots, G_k(z)$, then the random variable $Y = \sum_{i=1}^{k} X_i$ has PGF*

$$G(z) = \prod_{i=1}^{k} G_i(z).$$

Properties 22 *Let $X_1, \ldots, X_k$ be a sequence of independent random variables with the same PGF $G_X(z)$. If Y is a random variable with PGF $G_Y(z)$ and Y is independent of all X_i, then the random variable $S = X_1 + \ldots + X_Y$ has as PGF the function*

$$G_S(z) = G_Y(G_X(z)).$$

Definition 23 **Moment generating function** *(MGF) M_X of a random variable X is defined by*

$$M_X(z) = \mathbf{E}(e^{zX}).$$

Properties of moment generating functions:

Properties 24 *For a random variable X with MGF M_X it holds*

$$\mathbf{E}(X^k) = M_X^{(k)}(z)|_{z=0}.$$

IMPORTANT DISTRIBUTIONS

Two important distributions are connected with experiments called **Bernoulli trials**. They are experiments that have two possible outcomes:

- *success* with probability p
- *failure* with probability $q = 1 - p$

Coin tossing is an example of a Bernoulli trial.

Let values of a random variable X be the number of trials needed to obtain a success. Then

$$Pr(X = k) = q^{k-1}p$$

Such a probability distribution is called the **geometric distribution** and such a variable geometric random variable. It holds

$$\mathbf{E}X = \frac{1}{p} \qquad \mathbf{V}X = \frac{q}{p^2} \qquad G(z) = \frac{pz}{1 - qz}$$

Let now values of a random variable Y be the number of successes in n trials. Then

$$Pr(Y = k) = \binom{n}{k} p^k q^{n-k}$$

Such a probability distribution is called the **binomial distribution** and it holds

$$\mathbf{E}Y = np \qquad \mathbf{V}Y = npq \qquad G(z) = (q + pz)^n$$

and also

$$\mathbf{E}Y^2 = n(n-1)p^2 + np$$

BERNOULLI DISTRIBUTION

Let X be a binary random variable (called usually Bernouli or indicator random variable) that takes value 1 with probability p and 0 with probability $q = 1 - p$, then it holds

$$\mathbf{E}[X] = p \qquad \mathbf{V}X = pq \qquad G[z] = q + pz.$$

BINOMIAL DISTRIBUTION revisited

Let $X_1, \ldots, X_n$ be random variables having Bernoulli distribution with the common parameter p.
The random variable

$$X = X_1 + X_2 + \ldots + X_n$$

has binomial distribution denoted $B(n, p)$ with density function denoted

$$B(k, n, p) = \boldsymbol{Pr}(X = k) = \binom{n}{k} p^k q^{(n-k)}$$

POISSON DISTRIBUTION

Poisson distribution

Let $\lambda \in \mathbf{R}^{>0}$. The Poisson distribution with the parameter λ has the density function

$$p(x) = \begin{cases} \lambda^x \frac{e^{-\lambda}}{x!}, & \text{for } x = 0, 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases}$$

For large n the Poisson distribution is a good approximation to the Binomial distribution $B(n, \frac{\lambda}{n})$

Properties 25 *of a Poisson random variable X:*

$$\mathbf{E}[X] = \lambda \qquad \mathbf{V}X = \lambda \qquad G[z] = e^{\lambda(z-1)}$$

EXPECTATION+VARIANCE OF SUMS OF RANDOM VARIABLES

Let

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

where each X_i is a random variable which takes on value 1 (0) with probability p $(1 - p = q)$.
It clearly holds

$$\begin{aligned}
\mathbf{E}(X_i) &= p \\
\mathbf{E}(X_i^2) &= p \\
\mathbf{E}(S_n) &= \mathbf{E}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathbf{E}(X_i) = np \\
\mathbf{E}(S_n^2) &= \mathbf{E}((\sum_{i=1}^{n} X_i)^2) = \mathbf{E}(\sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} X_i X_j) = \\
&= \sum_{i=1}^{n} \mathbf{E}(X_i^2) + \sum_{i \neq j} \mathbf{E}(X_i X_j)
\end{aligned}$$

Hence

$$\begin{aligned}\mathbf{E}(S_n^2) &= \mathbf{E}((\sum_{i=1}^{n} X_i)^2) = \mathbf{E}(\sum_{i=1}^{n} X_i^2 + \sum_{i\neq j} X_i X_j) = \\ &= \sum_{i=1}^{n} \mathbf{E}(X_i^2) + \sum_{i\neq j} \mathbf{E}(X_i X_j)\end{aligned}$$

and therefore, if X_i, X_j are pairwise independent, as in this case, $\mathbf{E}(X_i X_j) =$ $= \mathbf{E}(X_i)\mathbf{E}(X_j)$ Hence

$$\begin{aligned}\mathbf{E}(S_n^2) &= np + 2\binom{n}{2}p^2 \\ &= np + n(n-1)p^2 \\ &= np(1-p) + n^2p^2 \\ &= n^2p^2 + npq \\ VAR[S_n] &= \mathbf{E}(S_n^2) - (\mathbf{E}(S_n))^2 = n^2p^2 + npq - n^2p^2 = npq\end{aligned}$$

MOMENT INEQUALITIES

Let X be a random variable that takes on values x with probability $p(x)$.

Theorem 26 *If* $\lambda > 0$, *then* k^{th} moment inequality *is*

$$Pr[|X| > \lambda] \leq \frac{\mathbf{E}(|X|^k)}{\lambda^k}$$

Proof

$$\begin{aligned} \mathbf{E}(|X|^k) &= \sum |x|^k p(x) \geq \sum_{|x|>\lambda} |x|^k p(x) \geq \\ &\geq \lambda^k \sum_{|x|>\lambda} p(x) = \lambda^k Pr[|X| > \lambda] \end{aligned}$$

Theorem 27 *If* $\lambda > 0$, *then* k^{th} moment inequality *is*

$$Pr[|X| > \lambda] \leq \frac{\mathbf{E}(|X|^k)}{\lambda^k}$$

Special cases

$$k \rightarrow 1 \qquad \lambda \rightarrow \lambda\mathbf{E}(|X|)$$

$$Pr[|X| \geq \lambda\mathbf{E}(|X|)] \leq \frac{1}{\lambda} \qquad \textbf{Markov's inequality}$$

$$k \rightarrow 2 \qquad X \rightarrow X - \mathbf{E}(X), \lambda \rightarrow \lambda\sqrt{V(X)}$$

$$Pr\left[|X - \mathbf{E}(X)| \geq \lambda\sqrt{V(X)}\right] \leq \frac{\mathbf{E}((X - \mathbf{E}(X))^2)}{\lambda^2 V(X)} =$$

$$= \frac{V(X)}{\lambda^2 V(X)} = \frac{1}{\lambda^2} \qquad \textbf{Chebyshev's inequality}$$

Another variant of Chebyshev's inequality:

$$\Pr[(X - \mathbf{E}(X)^2 \geq \lambda] \leq \frac{V(X)}{\lambda}$$

and this is one of the main reasons why variance is used.

Theorem 28 *If $g(x)$ is non-decreasing on $[0, \infty)$, then*

$$Pr[|X| > \lambda] \leq \frac{\mathbf{E}(g(X))}{g(\lambda)}$$

and taking $g(x) = e^{tx}$

$$Pr[|X| > \lambda] \leq \frac{\mathbf{E}(e^{tX})}{e^{t\lambda}} \qquad \textbf{Chernoff's inequality}$$

THE INCLUSION-EXCLUSION PRINCIPLE

Let $A_1, A_2, \ldots, A_n$ be events – not necessarily disjoint. The *Inclusion-Exclusion principle* states that

$$\begin{aligned} Pr\left[\bigcup_{i=1}^{n} A_i\right] &= \sum_{i=1}^{n} Pr(A_i) - \sum_{i<j} Pr(A_i \cap A_j) + \sum_{i<j<k} Pr(A_i \cap A_j \cap A_k) - \\ &- \ \ldots \ + (-1)^{k+1} \sum_{i_1<i_2<\ldots<i_k} Pr\left[\bigcap_{j=1}^{k} A_{i_j}\right] \ \ldots \ + \\ &+ (-1)^{n+1}\, Pr\left[\bigcap_{i=1}^{n} A_i\right] \end{aligned}$$

Bonferroni's inequalities

For every odd $k \leq n$

$$Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{j=1}^{k} (-1)^{j+1} \sum_{i_1<\ldots<i_j\leq n} Pr\left(\bigcap_{l=1}^{j} A_{i_l}\right)$$

For every even $k \leq n$

$$Pr\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{j=1}^{k} (-1)^{j+1} \sum_{i_1<\ldots<i_j\leq n} Pr\left(\bigcap_{l=1}^{j} A_{i_l}\right)$$

Special cases:

"Markov"-type inequality

$$Pr\left(\bigcup_{i} A_i\right) \leq \sum_{i} Pr(A_i)$$

"Chebyshev"-type inequality

$$Pr\left(\bigcup_{i} A_i\right) \geq \sum_{i} Pr(A_i) - \sum_{i<j} Pr(A_i \cap A_j)$$

CONDITIONAL EXPECTATION

It is natural to define conditional expectation of a random variable Y conditioned on an event E by

$$\mathbf{E}[Y|E] = \sum y Pr(Y = y|E).$$

Example Let we roll independently two perfect dice and let X_i be the number that shows on the ith dice and let X be sum of numbers on both dice.

$$\mathbf{E}[X|X_1 = 3] = \sum_x x Pr(X = x|X_1 = 3) = \sum_{x=4}^{9} x\frac{1}{6} = \frac{13}{2}$$

$$\mathbf{E}[X_1|X = 5] = \sum_{x=1}^{4} x Pr(X_1 = x|X = 5) = \sum_{x=1}^{4} x\frac{Pr(X_1 = x \cap X = 5)}{Pr(X = 5)} = \frac{5}{2}$$

For two random variables Y and Z, $\mathbf{E}[Y|Z]$ is defined to be a random variable $f(Z)$ that takes on the value $\mathbf{E}[Y|Z = z]$ when $Z = z$.

Theorem For any random variables Y, Z it holds

$$\mathbf{E}[Y] = \mathbf{E}[\mathbf{E}[Y|Z]].$$

SUMMATION with a STOPPING RULE

The following problem occurs in many applications:

For a given sequence $X_1, X_2, \ldots$ of random variables we need to compute

$$\mathbf{E}[\sum_{i=1}^{T} X_i],$$

where T is an integer-valued random variable.

So called Wald equation, on the next slide, shows how such expectation can be determined under special, but natural conditions. For that we define when T represents a stopping rule.

Definition Let $X_1, X_2, \ldots$ be a sequence of independent random variables, and T be an integer-valued random variable. T is called **stopping rule** relative to sequence $X_1, X_2, \ldots$, if the event $T = i$ is independent of $X_{i+1}, X_{i+2}, \ldots$.

The idea is that variables X_i are observed one after another, X_1, $X_2, \ldots$.
and
T represents the number of variables X_i observed when observation is finished.

Clearly, the event $T = i$ has to be independent of $X_{i+1}, X_{i+2}, \ldots$

EXAMPLES of STOPPING TIME

A stopping time corresponds to a strategy to stop a sequence of steps (say of gamblings) that is based only on the outcomes seen so far.

Examples of rules when a decision to stop gambling is a stopping time:

1. **First time the gambler wins 5 games in a row;**
2. **First time the gambler either wins or looses 1000 dolars;**

The rule "Last time the gambler wins 4 times in a row" is not a stopping time.

WALD's EQUATION

Theorem (Wald's equation) If $X_1, X_2, \ldots$ is a sequence of independent random variables that all have the same expectation μ and if T is a random variable, with a finite expectation, and T is a stopping rule for the sequence $\{X_i\}$, then

$$\mathbf{E}[\sum_{i=1}^{T} X_i] = \mu\mathbf{E}[T] = \mathbf{E}[X_i]\mathbf{E}[T].$$

Proof. Let us define an indicator variables $I_{i,T}$, for $i \geq 1$ as follow

$$I_{i,T} = \begin{cases} 1 & \textbf{if } i \leq T \\ 0 & \textbf{otherwise} \end{cases}$$

Clearly,

$$\sum_{i=1}^{T} X_i = \sum_{i=1}^{\infty} X_i I_{i,T}$$

and hence

$$\mathbf{E}[\sum_{i=1}^{T} X_i] = \mathbf{E}[\sum_{i=1}^{\infty} X_i \cdot I_{i,T}] = \sum_{i=1}^{\infty} \mathbf{E}[X_i \cdot I_{i,T}],$$

using countable additivity of expectations. ($\mathbf{E}[T] < \infty$ implies $\boldsymbol{Pr}[T = \infty] = 0$, and therefore $\sum_{i=1}^{\infty} \mathbf{E}[X_i \cdot I_{i,T}] < \infty$).

Since X_i are mutually independent, X_i and $I_{i,T}$ are also independent and we have

$$\mathbf{E}[\sum_{i=1}^{T} X_i] = \sum_{i=1}^{\infty} \mathbf{E}[X_i] \cdot \mathbf{E}[I_{i,T}] = \mu \sum_{i=1}^{\infty} \underline{\mathbf{E}}[I_{i,T}].$$

Theorem now follows from the observation that $T = \sum_{i=1}^{\infty} I_{i,T}$ and that

$$\mathbf{E}[T] = \sum_{i=1}^{\infty} \mathbf{E}[I_{i,T}].$$

EXAMPLE

- **Consider a gambling game in which a player first rolls one standard die. If the outcome of the roll is X, then he rolls X new standard dice and his gain Z is the sum of the outcomes of X dice,**
- **What is the expected gain of such a game?**
- **For $1 \le i \le X$, let Y_i be the outcome of the ith dice in the second round. Then**

$$\mathbf{E}[Z] = \mathbf{E}\left[\sum_{i=1}^{X} Y_i\right].$$

- **Since X is a stopping time, Wald's equality gives us**

$$\mathbf{E}[Z] = \mathbf{E}[X] \cdot \mathbf{E}[Y_i] = \left(\frac{7}{2}\right)^2 = \frac{49}{4}.$$

APPENDIX

APPENDIX

PUZLE - HOMEWORK

Puzzle 1 Given a biased coin, how to use it to simulate an unbiased coin?

Puzle 2 n people sit in a circle. Each person wears either red hat or a blue hat, chosen independently and uniformly at random. Each person can see the hats of all the other people, but not his/her hat. Based only upon what they see, each person votes on whether or not the total number of red hats is odd. Is there a scheme by which the outcome of the vote is correct with probability greater than 1/2.

BAYES' RULE STORY

- Reverend Thomas Bayes from England discovered the initial version of the "Bayes's law" around 1974, but soon stopped to believe in it.
- In behind were two philosophical questions
  - Can an effect determine its cause?
  - Can we determine the existence of God by observing nature?
- Bayes law was not written for long time as formula, only as the statement: By updating our initial belief about something with objective new information, we can get a new and improved belief.
- Bayes used a tricky thought experiment to demonstrate his law.

- Bayes' rule was later invented independently by Pierre Simon Laplace, perhaps the greatest scientist of 18th century, but at the end he also abounded it.
- Till the 20 century theoreticians considered Bayes rule as unscientific. Bayes rule had for centuries several proponents and many opponents in spite that it has turned out to be very useful in practice.
- Bayes rule was used to help to create rules of insurance industries, to develop strategy for artillery during the first and even Second World War (and also a great Russian mathematician Kolmogorov helped to develop it for this purpose).
- It was used much to decrypt ENIGMA codes during 2WW, due to Turing, and also to locate German submarines.