

Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 04 - Discrete random variables

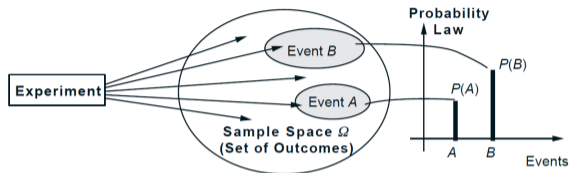
Salvatore Ruggieri

Department of Computer Science

University of Pisa, Italy

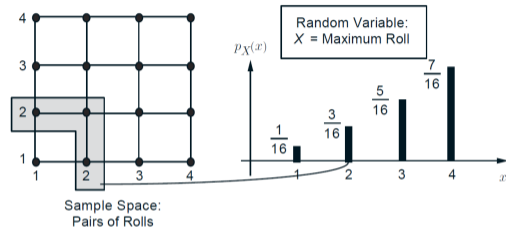
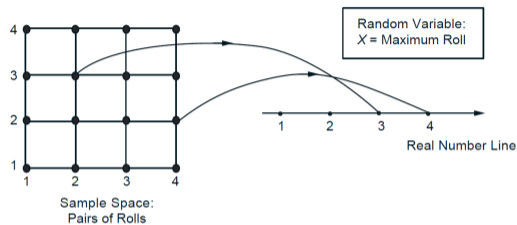
salvatore.ruggieri@unipi.it

Experiments



- **Experiment:** roll two independent 4 sided die.
- We are interested in probability of the *maximum of the two rolls*.
- Modeling so far
 - ▶ $\Omega = \{1, 2, 3, 4\} \times \{1, 2, 3, 4\} = \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 1), \dots, (4, 4)\}$
 - ▶ $A = \{\text{maximum roll is 2}\} = \{(1, 2), (2, 1), (2, 2)\}$
 - ▶ $P(A) = P(\{(1, 2), (2, 1), (2, 2)\}) = 3/16$

Random variables

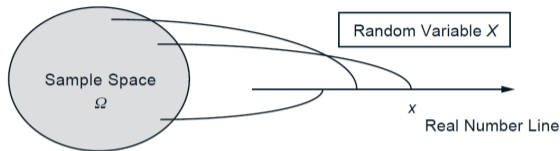


- Modeling $X : \Omega \rightarrow \mathbb{R}$

- ▶ $X((a, b)) = \max(a, b)$
- ▶ $A = \{\text{maximum roll is 2}\} = \{(a, b) \in \Omega \mid X((a, b)) = 2\} = X^{-1}(2)$
- ▶ $P(A) = P(X^{-1}(2)) = \frac{3}{16}$
- ▶ We write $P_X(X = 2) \stackrel{\text{def}}{=} P(X^{-1}(2))$

Induced probability

(Discrete) Random variables



- A random variable is a function $X : \Omega \rightarrow \mathbb{R}$
 - ▶ it transforms Ω into a more tangible sample space \mathbb{R}
 - from (a, b) to $\min(a, b)$
 - ▶ it decouples the details of a specific Ω from the probability of events of interest
 - from $\Omega = \{H, T\}$ or $\Omega = \{\text{good}, \text{bad}\}$ or $\Omega = \dots$ to $\{0, 1\}$
 - ▶ it is not 'random' nor 'variable'

DEFINITION. Let Ω be a sample space. A *discrete random variable* is a function $X : \Omega \rightarrow \mathbb{R}$ that takes on a finite number of values a_1, a_2, \dots, a_n or an infinite number of values a_1, a_2, \dots

Probability Mass Function (PMF)

DEFINITION. The *probability mass function* p of a discrete random variable X is the function $p : \mathbb{R} \rightarrow [0, 1]$, defined by

$$p(a) = P(X = a) \quad \text{for } -\infty < a < \infty.$$

- Support of X is $\{a \in \mathbb{R} \mid P(X = a) > 0\} = \{a_1, a_2, \dots\}$
 - ▶ $p(a_i) > 0$ for $i = 1, 2, \dots$
 - ▶ $p(a_1) + p(a_2) + \dots = 1$
 - ▶ $p(a) = 0$ if $a \notin \{a_1, a_2, \dots\}$

Cumulative Distribution Function (CDF) and CCDF

DEFINITION. The *distribution function* F of a random variable X is the function $F : \mathbb{R} \rightarrow [0, 1]$, defined by

$$F(a) = P(X \leq a) \quad \text{for } -\infty < a < \infty.$$

- $F(a) = P(X \in \{a_i \mid a_i \leq a\}) = P(X \leq a) = \sum_{a_i \leq a} p(a_i)$
- if $a \leq b$ then $F(a) \leq F(b)$
- $P(a < X \leq b) = F(b) - F(a) = \sum_{a < a_i \leq b} p(a_i)$

[Non-decreasing]

Complementary cumulative distribution function (CCDF)

$$\bar{F}(a) = P(X > a) = 1 - P(X \leq a) = 1 - F(a)$$

- $\bar{F}(a) = P(X \in \{a_i \mid a_i > a\}) = P(X > a) = \sum_{a_i > a} p(a_i)$

See R script

$$X \sim U(m, M)$$

Uniform discrete distribution

A discrete random variable X has the *uniform distribution* with parameters $m, M \in \mathbb{Z}$ such that $m \leq M$, if its pmf is given by

$$p(a) = \frac{1}{M - m + 1} \quad \text{for } a = m, m + 1, \dots, M$$

We denote this distribution by $U(m, M)$.

- **Intuition:** all integers in $[m, M]$ have equal chances of being observed.

$$F(a) = \frac{\lfloor a \rfloor - m + 1}{M - m + 1} \quad \text{for } m \leq a \leq M$$

See R script

Benford's law

A discrete random variable X has the *Benford's distribution*, if its pmf is given by

$$p(a) = \log_{10} \left(1 + \frac{1}{a} \right) \quad \text{for } a = 1, 2, \dots, 9$$

We denote this distribution by *Ben*.

- Plausible and empirically adequate model for to the frequency distribution of leading digits in many real-life numerical datasets.
- See [Wikipedia](#) for its interesting history and applications!

See R script

$X \sim \text{Ber}(p)$

DEFINITION. A discrete random variable X has a **Bernoulli distribution** with parameter p , where $0 \leq p \leq 1$, if its probability mass function is given by

$$p_X(1) = P(X = 1) = p \quad \text{and} \quad p_X(0) = P(X = 0) = 1 - p.$$

We denote this distribution by $\text{Ber}(p)$.

- X models success/failure in tossing a coin (H, T), testing for a disease (infected, not infected), membership in a set (member, non-member), etc.
- p_X is the *pmf* (to distinguish from parameter p)
- Alternative definition: $p_X(a) = p^a \cdot (1 - p)^{1-a}$ for $a \in \{0, 1\}$

See R script

Identically distributed random variables

Two random variables X and Y are said *identically distributed* (in symbols, $X \sim Y$), if $F_X = F_Y$, i.e.,

$$F_X(a) = F_Y(a) \quad \text{for } a \in \mathbb{R}$$

- Identically distributed does **not** mean equal
- Toss a fair coin
 - ▶ let X be 1 for H and 0 for T
 - ▶ let Y be $1 - X$
- $X \sim \text{Ber}(0.5)$ and $Y \sim \text{Ber}(0.5)$
- Thus, $X \sim Y$ but are clearly always different.

Joint p.m.f.

- For a same Ω , several random variables can be defined
 - ▶ Random variables related to the same experiment often influence one another
 - ▶ $\Omega = \{(i, j) \mid i, j \in 1, \dots, 6\}$ rolls of two dies
 - $X((i, j)) = i + j$ and $Y((i, j)) = \max(i, j)$
 - $P(X = 4, Y = 3) = P(X^{-1}(4) \cap Y^{-1}(3)) = P(\{(3, 1), (1, 3)\}) = 2/36$
 - ▶ $\Omega = \{f, m\} \times \mathbb{N} \times \{+, -\}$ (testing for Covid-19 - multivariate)
 - $G((g, a, c)) = 0$ if $g = f$ and 1 otherwise $A((g, a, c)) = a$
 - $Y((g, a, c)) = 0$ if $c = -$ and 1 otherwise
- In general:

$$P_{XY}(X = a, Y = b) = P(\{\omega \in \Omega \mid X(\omega) = a \text{ and } Y(\omega) = b\}) = P(X^{-1}(a) \cap Y^{-1}(b))$$

DEFINITION. The *joint probability mass function* p of two discrete random variables X and Y is the function $p : \mathbb{R}^2 \rightarrow [0, 1]$, defined by

$$p(a, b) = P(X = a, Y = b) \quad \text{for } -\infty < a, b < \infty.$$

Joint and marginal p.m.f.

- **Joint distribution function** $F : \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$:

$$F_{XY}(a, b) = P(X \leq a, Y \leq b) = \sum_{a_i \leq a, b_i \leq b} p(a_i, b_i)$$

- By generalized additivity, the **marginal p.m.f.**'s can be derived: [Tabular method]

$$p_X(a) = P_X(X = a) = \sum_b P_{XY}(X = a, Y = b) \quad p_Y(b) = P_Y(Y = b) = \sum_a P_{XY}(X = a, Y = b)$$

and the marginal distribution function of X as:

$$F_X(a) = P_X(X \leq a) = \lim_{b \rightarrow \infty} F_{XY}(a, b) \quad F_Y(b) = P_Y(Y \leq b) = \lim_{a \rightarrow \infty} F_{XY}(a, b)$$

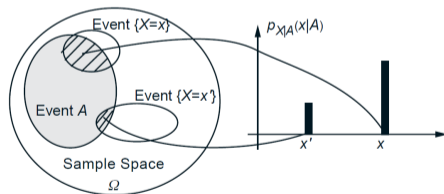
- Deriving the joint p.m.f. from marginal p.m.f.'s is not always possible!
 - ▶ **Exercise at home.** Prove it (hint: find two joint p.m.f.'s with the same marginals)
- Deriving the joint p.m.f. from marginal p.m.f.'s is possible for independent events!
 - ▶ $\Omega = \{1, 2, 3, 4\} \times \{1, 2, 3, 4\}$, $X((a, b)) = a$, $Y((a, b)) = b$
 - ▶ $P(X = 1, Y = 2) = 1/16 = 1/4 \cdot 1/4 = P(X = 1) \cdot P(Y = 2)$

Conditional distribution

Conditional distribution

Consider the joint distribution P_{XY} of X and Y . The conditional distribution of X given $Y \in B$ with $P_Y(Y \in B) > 0$, is the function $F_{X|Y \in B} : \mathbb{R} \rightarrow [0, 1]$:

$$F_{X|Y \in B}(a) = P_{X|Y}(X \leq a | Y \in B) = \frac{P_{XY}(X \leq a, Y \in B)}{P_Y(Y \in B)} \quad \text{for } -\infty < a < \infty$$



- Distribution of X after knowing $Y \in B$.
- Chain rule: $P_{XY}(X \leq a, Y \in B) = P_{X|Y}(X \leq a | Y \in B)P_Y(Y \in B)$
- What if the distribution does not change w.r.t. the prior P_X ?

(Machine Learning) Binary Classifiers

- $\Omega = \{f, m\} \times \mathbb{N} \times \{+, -\}$
- Predictive Features and True-Class as Random Variables:
 - ▶ gender: $G((g, a, c)) = 1$ if g is f and 0 otherwise
 - ▶ age: $A((g, a, c)) = a$
 - ▶ has-covid: $Y((g, a, c)) = 1$ if $c = +$ and 0 otherwise
- Binary Classifier as a Random Variable:
 - ▶ $\hat{Y}((g, a, c)) = 1$ if $clf((g, a)) = +$ and 0 otherwise
where $clf : \{f, m\} \times \mathbb{N} \rightarrow \{+, -\}$ is a function over predictive features
- $P(Y = \hat{Y})$, i.e., $P(\{\omega \in \Omega \mid Y(\omega) = \hat{Y}(\omega)\})$ *[True Accuracy]*
- $P(Y = 1 \mid \hat{Y} = 1)$ *[True Precision]*
- $P(\hat{Y} = 1 \mid Y = 1)$ *[True Recall]*
- **Such probabilities are unknown!** They can only be estimated on a sample (*test set*)

Independence of two random variables

Independence $X \perp\!\!\!\perp Y$

A random variable X is independent from a random variable Y , if for all $P_Y(Y \leq b) > 0$:

$$P_{X|Y}(X \leq a | Y \leq b) = P_X(X \leq a) \quad \text{for } -\infty < a < \infty$$

- Properties
 - ▶ $X \perp\!\!\!\perp Y$ iff $P_{XY}(X \leq a, Y \leq b) = P_X(X \leq a) \cdot P_Y(Y \leq b)$ for $-\infty < a, b < \infty$
 - ▶ $X \perp\!\!\!\perp Y$ iff $Y \perp\!\!\!\perp X$ *[Symmetry]*
- For X, Y **discrete** random variables:
 - ▶ $X \perp\!\!\!\perp Y$ iff $P_{XY}(X = a, Y = b) = P_X(X = a) \cdot P_Y(Y = b)$ for $-\infty < a, b < \infty$
 - ▶ **Exercise at home.** Prove it!
 - ▶ $X \perp\!\!\!\perp Y$ iff $P_{XY}(X \in \mathcal{A}, Y \in \mathcal{B}) = P_X(X \in \mathcal{A}) \cdot P_Y(Y \in \mathcal{B})$ for $\mathcal{A}, \mathcal{B} \subseteq \mathbb{R}$
 - ▶ **Exercise at home.** Prove it!

See R script

Sum of independent discrete random variables

ADDING TWO INDEPENDENT DISCRETE RANDOM VARIABLES. Let X and Y be two independent discrete random variables, with probability mass functions p_X and p_Y . Then the probability mass function p_Z of $Z = X + Y$ satisfies

$$p_Z(c) = \sum_j p_X(c - b_j)p_Y(b_j),$$

where the sum runs over all possible values b_j of Y .

- **Proof (sketch).** $P(Z = c) = \sum_j P(Z = c | Y = b_j) \cdot P(Y = b_j) = \sum_j P(X = c - b_j | Y = b_j) \cdot P(Y = b_j) = \sum_j P(X = c - b_j)P(Y = b_j)$

Independence of multiple random variables

Independence (factorization formula)

Random variables X_1, \dots, X_n are independent, if:

$$P_{X_1, \dots, X_n}(X_1 \leq a_1, \dots, X_n \leq a_n) = \prod_{i=1}^n P_{X_i}(X_i \leq a_i) \quad \text{for } -\infty < a_1, \dots, a_n < \infty$$

- X_1, \dots, X_n **discrete** random variables are independent iff:

$$P_{X_1, \dots, X_n}(X_1 = a_1, \dots, X_n = a_n) = \prod_{i=1}^n P_{X_i}(X_i = a_i) \quad \text{for } -\infty < a_1, \dots, a_n < \infty$$

- **Definition:** X_1, \dots, X_n are **i.i.d.** (independent and identically distributed) if X_1, \dots, X_n are independent and $X_i \sim F$ for $i = 1, \dots, n$ for some distribution F

$X \sim \text{Bin}(n, p)$

DEFINITION. A discrete random variable X has a *binomial distribution* with parameters n and p , where $n = 1, 2, \dots$ and $0 \leq p \leq 1$, if its probability mass function is given by

$$p_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, \dots, n.$$

We denote this distribution by $\text{Bin}(n, p)$.

- X models the number of successes in n Bernoulli trials (How many H's when tossing n coins?)
- **Intuition:** for X_1, X_2, \dots, X_n such that $X_i \sim \text{Ber}(p)$ and independent (**i.i.d.**):

$$X = \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$$

- $p^k \cdot (1-p)^{n-k}$ is the probability of observing first k H's and then $n-k$ T's
- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ number of ways to choose the first k variables *[Binomial coefficient]*
- $p_X(k)$ computationally expensive to calculate (no closed formula, but approximation/bounds)
- **Exercise at home.** Prove $X_1 + X_2 \sim \text{Bin}(2, p)$ using the sum of independent random variables.

See R script

$$X \sim \text{Bin}(n, p)$$

DEFINITION. A discrete random variable X has a **binomial distribution** with parameters n and p , where $n = 1, 2, \dots$ and $0 \leq p \leq 1$, if its probability mass function is given by

$$p_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, \dots, n.$$

We denote this distribution by **$\text{Bin}(n, p)$** .

- **Exercise:** there are c bikes shared among n persons. Assuming that each person needs a bike with probability p , what is the probability that all bikes will be in use?

$$P(X = c) = \binom{n}{c} p^c \cdot (1-p)^{n-c} = \text{dbinom}(c-1, n, p)$$

$X \sim \text{Geo}(p)$

DEFINITION. A discrete random variable X has a *geometric distribution* with parameter p , where $0 < p \leq 1$, if its probability mass function is given by

$$p_X(k) = P(X = k) = (1 - p)^{k-1} p \quad \text{for } k = 1, 2, \dots$$

We denote this distribution by $\text{Geo}(p)$.

- X models the number of Bernoulli trials before a success (how many tosses to have a H?)
- **Intuition:** for X_1, X_2, \dots such that $X_i \sim \text{Ber}(p)$ i.i.d.:

$$X = \min_i (X_i = 1) \sim \text{Geo}(p)$$

- $\bar{F}(a) = P(X > a) = (1 - p)^{\lfloor a \rfloor}$
- $F(a) = P(X \leq a) = 1 - \bar{F}(a) = 1 - (1 - p)^{\lfloor a \rfloor}$

See R script

You cannot always loose

- H is 1, T is 0, $0 < p < 1$
- $B_n = \{\text{T in the first } n\text{-th coin tosses}\}$
- $P(\cap_{n \geq 1} B_i) = ?$
- $X \sim \text{Geom}(p)$
- $P(B_n) = P(X > n) = (1 - p)^n$
- $P(\cap_{n \geq 1} B_n) = \lim_{n \rightarrow \infty} P(B_n) = \lim_{n \rightarrow \infty} (1 - p)^n = 0$
- $P(\cap_{n \geq 1} B_n) = \lim_{n \rightarrow \infty} P(B_n)$ for B_n non-increasing

[σ -additivity, see Lesson 01]

But if you lost so far, you can lose again

Memoryless property

For $X \sim \text{Geo}(p)$, and $n, k = 0, 1, 2, \dots$

$$P(X > n + k | X > k) = P(X > n)$$

Proof

$$\begin{aligned} P(X > n + k | X > k) &= \frac{P(\{X > n + k\} \cap \{X > k\})}{P(\{X > k\})} \\ &= \frac{P(\{X > n + k\})}{P(\{X > k\})} \\ &= \frac{(1 - p)^{n+k}}{(1 - p)^k} \\ &= (1 - p)^n = P(X > n) \end{aligned}$$

$$X \sim NBin(n, p)$$

Negative binomial (or Pascal distribution)

A discrete random variable X has a negative binomial with parameters n and p , where $n = 0, 1, 2, \dots$ and $0 < p \leq 1$, if its probability mass function is given by

$$p_X(k) = P(X = k) = \binom{k+n-1}{k} (1-p)^k \cdot p^n \quad \text{for } k = 0, 1, 2, \dots$$

- X models the number of failures before the n -th success in Bernoulli trials (how many T's to have n H's?)
- **Intuition:** for X_1, X_2, \dots, X_n such that $X_i \sim Geo(p)$ i.i.d.:

$$X = \sum_{i=1}^n X_i - n \sim NBin(n, p)$$

- $(1-p)^k \cdot p^n$ is the probability of observing first k T's and then n H's
- $\binom{k+n-1}{k} = \frac{(k+n-1)!}{k!(n-1)!}$ number of ways to choose the first k variables among $k+n-1$ (the last one must be a success!)

See R script

$X \sim Poi(\mu)$

DEFINITION. A discrete random variable X has a *Poisson distribution* with parameter μ , where $\mu > 0$ if its probability mass function p is given by

$$p(k) = P(X = k) = \frac{\mu^k}{k!} e^{-\mu} \quad \text{for } k = 0, 1, 2, \dots$$

We denote this distribution by $Pois(\mu)$.

- X models the number of events in a fixed interval if these events occur with a known constant mean rate μ and independently of the last event
 - ▶ telephone calls arriving in a system
 - ▶ number of patients arriving at an hospital
 - ▶ customers arriving at a counter
- μ denotes the mean number of events
- $Bin(n, \mu/n)$ is the number of successes in n trials, assuming $p = \mu/n$, i.e., $p \cdot n = \mu$
- When $n \rightarrow \infty$: $Bin(n, \mu/n) \rightarrow Poi(\mu)$ [Law of rare events]
 - ▶ Number of typos in a book, number of cars involved in accidents, etc.

See R script

The discrete Bayes' rule

BAYES' RULE. Suppose the events C_1, C_2, \dots, C_m are disjoint and $C_1 \cup C_2 \cup \dots \cup C_m = \Omega$. The conditional probability of C_i , given an arbitrary event A , can be expressed as:

$$P(C_i | A) = \frac{P(A | C_i) \cdot P(C_i)}{P(A | C_1)P(C_1) + P(A | C_2)P(C_2) + \dots + P(A | C_m)P(C_m)}.$$

- **Definition.** Conditional p.m.f. of X given $Y = b$ with $P_Y(Y = b) > 0$


$$p_{X|Y}(a|b) = \frac{p_{XY}(a, b)}{p_Y(b)} \quad \text{i.e.,} \quad P_{X|Y}(X = a | Y = b) = \frac{P_{XY}(X = a, Y = b)}{P_Y(Y = b)}$$

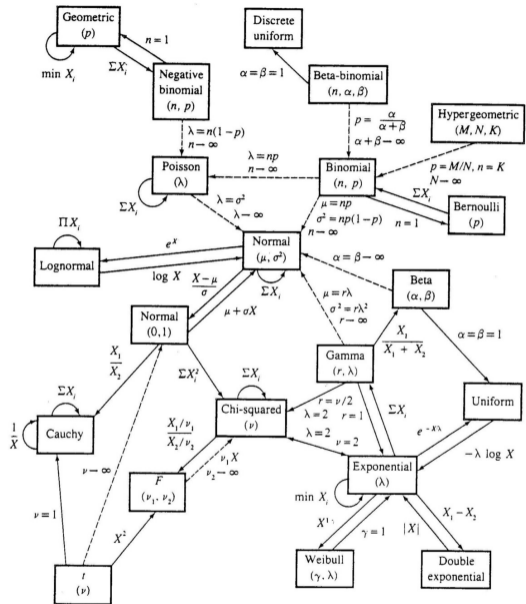
- Discrete Bayes' rule:

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)} = \frac{p_{Y|X}(y|x)p_X(x)}{\sum_{a \in \text{dom}(X)} p_{Y|X}(y|a)p_X(a)}$$

- **Exercise at home.** A machine fails after n days with a p.m.f. $X \sim \text{Geo}(p)$. p is known to be either $p = 0.1$ or 0.05 with equal probability. What can we say about the distribution of p given n ? Code your solution in R.

Common distributions

- Probability distributions at Wikipedia
- Probability distributions in R
-  C. Forbes, M. Evans, N. Hastings, B. Peacock (2010) Statistical Distributions, 4th Edition Wiley



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).