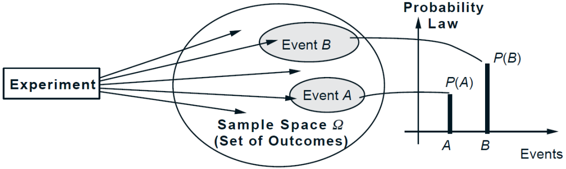
Statistical Methods for Data Science

Lesson 03 - Discrete random variables

Salvatore Ruggieri

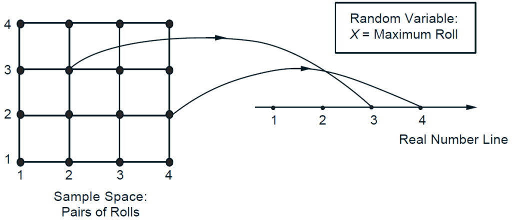
Department of Computer Science
University of Pisa
salvatore.ruggieri@unipi.it

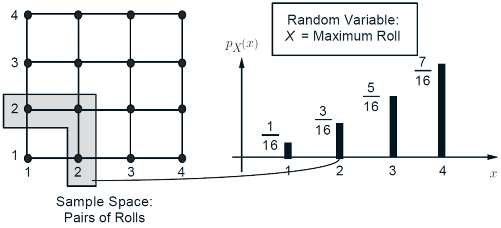
Experiments



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

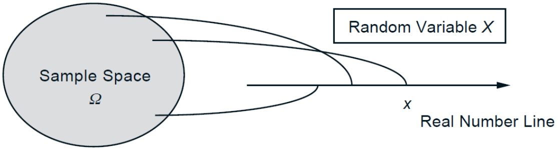
Random variables





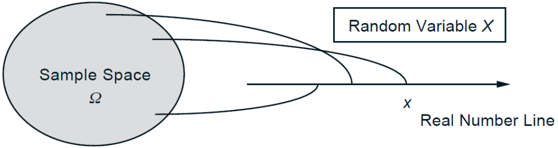
- Modeling $X : \Omega \to \mathbb{R}$
  - $X((a,b)) = max(a,b)$
  - $A = \{\text{maximum roll is } 2\} = \{(a,b) \in \Omega \mid X((a,b)) = 2\} = X^{-1}(2)$
  - $P(A) = P(X^{-1}(2)) = 3/16$
  - We write $P_X(X = 2) \stackrel{\text{def}}{=} P(X^{-1}(2))$ *[Induced probability]*

(Discrete) Random variables



- A random variable is a function $X : \Omega \to \mathbb{R}$
  - 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 $\Omega = \{\text{H, T}\}$ or $\Omega = \{\text{good, bad}\}$ or $\Omega = \ldots$ to $\{0, 1\}$

(Discrete) Random variables



- A random variable is a function $X : \Omega \to \mathbb{R}$
  - 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 $\Omega = \{\text{H, T}\}$ or $\Omega = \{\text{good, bad}\}$ or $\Omega = \ldots$ to $\{0, 1\}$

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

Probability Mass Function (PMF)

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

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

- Sample space $\mathbb{R}$ but support is $\{a_1, \ldots, a_n\}$
  - $p(a_i) > 0$ for $i = 1, 2, \ldots$
  - $p(a_1) + p(a_2) + \ldots = 1$
  - $p(a) = 0$ if $a \notin \{a_1, a_2, \ldots\}$
- "$X = a$" shorthand for the event $\{a\} \subseteq \mathbb{R}$

Cumulative Distribution Function (CDF) and CCDF

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

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

- $F(a) = P(\{a_i \mid a_i \leq a\}) = \sum_{a_i \leq a} p(a_i)$
- if $a \leq b$ then $F(a) \leq F(b)$ *[Non-decreasing]*
- $P(a < X \leq b) = F(b) - F(a) = \sum_{a < a_i \leq b} p(a_i)$

Cumulative Distribution Function (CDF) and CCDF

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

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

- $F(a) = P(\{a_i \mid a_i \leq a\}) = \sum_{a_i \leq a} p(a_i)$
- if $a \leq b$ then $F(a) \leq F(b)$ *[Non-decreasing]*
- $P(a < X \leq b) = F(b) - F(a) = \sum_{a < a_i \leq b} p(a_i)$

Complementary cumulative distribution function (CCDF)

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

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

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

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

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

$X \sim Ben$

Benford's law

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

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

We denote this distribution by *Ben*.

- Related to the frequency distribution of leading digits in many real-life numerical datasets.
- See **Wikipedia** for its interesting history!

See R script

$X \sim Ber(p)$

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

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

We denote this distribution by $Ber(p)$.

- 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)
- Also, $p_X(a) = p^a \cdot (1-p)^{1-a}$ for $a \in \{0, 1\}$

See R script

$X \sim Bin(n, p)$

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

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

We denote this distribution by $Bin(n, p)$.

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

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

- $p^k \cdot (1-p)^{n-k}$ 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
- $p_X(k)$ computationally expensive to calculate (no closed formula, but approximation/bounds)

See R script

i.d. random variables

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 n times, where n is odd
  - let X be the number of heads
  - let Y be the number of tails
- $X \sim Bin(n, 0.5)$ and $Y \sim Bin(n, 1 - 0.5) = Bin(n, 0.5)$
- Thus, $X \sim Y$ but are clearly always unequal.

$X \sim Geo(p)$

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

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

We denote this distribution by $Geo(p)$.

- X models the number of trials before a success (how many tosses to have a H?)
- **Intuition**: for $X_1, X_2, \ldots$ such that $X_i \sim Ber(p)$ (**and independent**):

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

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

See R script

You cannot always loose

- H is 1, T is 0, $0 < p < 1$
- $B_n = \{$T in the first n-th coin tosses$\}$
- $P(\cap_{n\geq 1} B_i) = \ ?$

You cannot always loose

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

You cannot always loose

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

But if you lost so far, you can lose again

Memoryless property

For $X \sim Geo(p)$, and $n, k = 0, 1, 2, \ldots$

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

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

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 (how many T's to have n H's?)
- **Intuition**: for $X_1, X_2, \dots, X_n$ such that $X_i \sim Geo(p)$ (**and independent**):

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

- $(1-p)^k \cdot p^n$ 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

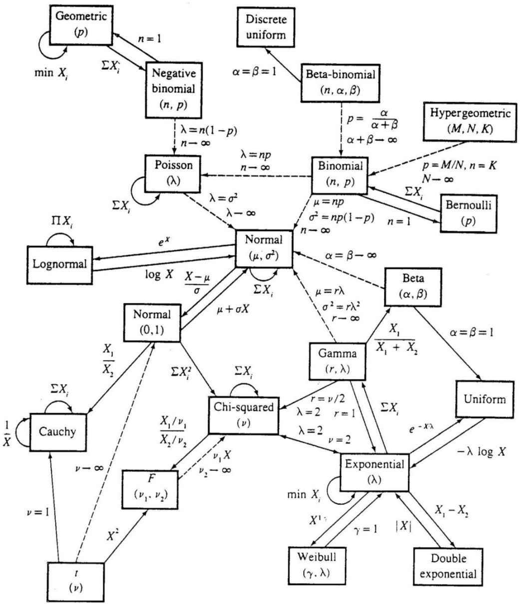
$$p(k) = P(X = k) = \frac{\mu^k}{k!} e^{-\mu} \quad \text{for } k = 0, 1, 2, \ldots.$$

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 $n \to \infty$: $Bin(n, \mu/n) \to Poi(\mu)$ *[Law of rare events]*

See R script

Common distributions



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