# *Introduction to Randomized Algorithms II*

Joaquim Madeira

Version 0.4 – October 2024

# Overview

- Discrete Probability
- Statistical Experiments and Events
- Probabilities and Random Variables
- Application Examples and Problems

# DISCRETE PROBABILITY

# Discrete Probability

- Chance enters into many attempts to understand the world we live in
- A theory of probability allows us to calculate the likelihood of complex events
- Probabilities are called "discrete" if we can compute the probabilities of all events by summation

# Probability Space

- Probability theory starts with the idea of a probability space $(\Omega, Pr)$
  - A set $\Omega$ of all things that can happen
  - A rule assigning a probability $Pr(\omega)$ to each elementary event $\omega$ in $\Omega$

# Probability Distribution

- For a discrete probability space
  - $\Pr(\omega) \geq 0$
  - $\sum \Pr(\omega) = 1$
- Pr is the probability distribution
  - It distributes the total probability among the elementary events

# Example – Fair dice

- Roll one fair 6-sided die

D = { ⚀, ⚁, ⚂, ⚃, ⚄, ⚅ }

  - Each of the 6 possibilities has probability 1/6

- Roll a pair of fair dice
  - Set of elementary events : $D^2$ = ?
  - Probability of each event ?

# Example – "Loaded" dice

- Distribution of probabilities

$$\Pr_1(\text{⚀}) = \Pr_1(\text{⚅}) = \frac{1}{4};$$

$$\Pr_1(\text{⚁}) = \Pr_1(\text{⚂}) = \Pr_1(\text{⚃}) = \Pr_1(\text{⚄}) = \frac{1}{8}$$

- Probability of each event in $D^2$ ?

$$\Pr_{11}(d\,d') = \Pr_1(d)\,\Pr_1(d')$$

# Example – Fair die + "Loaded" die

- Consider the case of one fair die and one loaded die

$$\Pr_{01}(d\,d') = \text{Pro}(d)\ \Pr_1(d'), \qquad \text{where } \text{Pro}(d) = \tfrac{1}{6};$$

- Real-world dice do not turn up equally often on each side !!
  - No perfect symmetry !!
- BUT, 1/6 is usually close to the truth…

# Example – Doubles are thrown

- The event that "doubles are thrown"

{⚀⚀, ⚁⚁, ⚂⚂, ⚃⚃, ⚄⚄, ⚅⚅}

- Probability of an event A

$$\Pr(\omega \in A) = \sum_{\omega \in A} \Pr(\omega)$$

- Pr("doubles are thrown") = ?
  - When is this event more probable ?

# Statistical experiments and events

- Statistical experiment
  - Repeatable experiment where the particular outcome of a trial cannot be predicted with certainty
- Sample space, *S*
  - Set with the representation of all possible outcomes of an experiment
  - I.e., set of elementary events
- Examples ?

# Statistical experiments and events

- Event, *E*
  - A set of elementary events
  - Any subset of a sample space

- Examples
  - {2, 4, 6} – Getting an even number when throwing a 6-sided die
    - Probability ?
  - …

# Probabilities

- The probability of an event, *E*, describes the degree of uncertainty of that event

- $0 \leq P[E] \leq 1$
- $P[S] = 1$
- $P[\emptyset] = 0$

- $P[A \cup B] = P[A] + P[B]$, if *A* and *B* are disjoint

- If $(A \cap B) \neq \emptyset$, $P[A \cup B] = P[A] + P[B] - P[A \cap B]$

# PROBABILITY DISTRIBUTIONS

# Uniform probability distribution on S

- Each elementary event in $S$ has the same probability

$$P[\{A\}] = P[\{B\}] = \frac{1}{|S|}$$

# Simple problem

- Throwing a 6-sided fair die
- What is the probability of getting an even number ?
- What is the probability of getting a number larger than 2 ?
- What is the probability of getting an even number <u>or</u> a number larger than 2 ?

# Another simple problem

- Tossing three fair coins
- What is the sample space $S_3$ ?
- How high is the probability of getting at least one head ?
- And at least two heads ?
- Idea: relate to the binary representation
- Idea: triangular representation – paths

# More difficult problem

- Tossing $n$ fair coins
- How large is the probability to get "head" exactly k times ?
- How large is the probability to get "head" at least k times ?
- You can use your code to check your answers…

# Binomial probability distribution

- Characterizes the probability of obtaining k "successes" in n experiments

$$S = \{0,1,2,\ldots,n\}$$

$$P[X=k] = \binom{n}{k} p^k (1-p)^{n-k}, k = 0,1,2,\ldots,n$$

- Check your previous answers !!

# Tasks

- What is the probability of getting 6 heads in 15 tosses of a fair coin ?
  - Estimate the value with simulated experiments
  - Check that you got the correct value by computing the probability from the binomial distribution

- Now, consider that $P[\text{heads}] = 2 \times P[\text{tails}]$

# Independent events

- Two events *A* and *B* are said to be independent, if the occurrence of one does not affect the occurrence of the other

- Two events *A* and *B* are independent, if

$$P[A \text{ and } B] = P[A] \times P[B]$$

# Conditional probability

- Conditional probability of event A given event B

$$P[A|B] = \frac{P[A \text{ and } B]}{P[B]}, P[B] \neq 0$$

  - What happens if they are independent ?

- Bayes' Theorem

$$P[A|B] = \frac{P[B|A]\, P[A]}{P[B]}$$

# RANDOM VARIABLES

# Random variable

- A random variable is a function that assigns a real number to each elementary event of the sample space
- Discrete r. v. – countable set of real values
  - Values obtained by throwing a dice numbered 1 to 6
  - Assigning 0 to tail and 1 to heads when tossing a coin
- Continuous r. v. – interval or collection of intervals
  - Examples: temperature of a room or weight of product

# Example – Throwing two dice

- S(w) = sum of spots on the dice roll w
- What is the probability that the spots total 7 ?

Pr(⚀⚅) + Pr(⚁⚄) + Pr(⚂⚃)

+ Pr(⚃⚂) + Pr(⚄⚁) + Pr(⚅⚀)

- Fair dice ?
- Loaded dice ?

# Example – Throwing two dice

- A random variable is characterized by the probability distribution of its values

| s | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $P_{r00}[S=s]$ | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |
| $P_{r11}[S=s]$ | $\frac{4}{64}$ | $\frac{4}{64}$ | $\frac{5}{64}$ | $\frac{6}{64}$ | $\frac{7}{64}$ | $\frac{12}{64}$ | $\frac{7}{64}$ | $\frac{6}{64}$ | $\frac{5}{64}$ | $\frac{4}{64}$ | $\frac{4}{64}$ |

# Sequence of numbers – Average value

- Mean
  - Sum of all values, divided by the number of values
- Median
  - Middle value, numerically
- Mode
  - Value that occurs most often

# statistics – Python module

- Computing mathematical statistics of numeric data
- Averages and measures of central location
  - mean(…)
  - median(…)
  - mode(…)
  - …
- Measures of spread
  - stdev(…)
  - variance(…)
  - …

# Discrete random variable – Features

- Mean – Expected value

$$\mu = E[X] = \sum_{n} X_n P[X = X_n]$$

- Variance
  - Measures how far a set of numbers are spread out

$$\sigma^2 = E[(X - \mu)^2] = E[X^2] - \mu^2$$

- $\sigma$ is the standard deviation

# Sum of independent random vars.

- Let $Z = X + Y$ be the sum of two independent random variables, defined on the same probability space

$$E[Z] = E[X + Y] = E[X] + E[Y]$$

$$\sigma_Z^2 = E[(X + Y - \mu)^2] = \sigma_X^2 + \sigma_Y^2$$

# Sum of independent random vars.

- Let $S_n$ be the sum of $n$ independent and identically distributed random variables

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

$$E[S_n] = n \times E[X]$$

$$\sigma_{S_n}^2 = n \times \sigma_X^2$$

# Estimating the mean of a rand. var.

- Set of independent empirical observations
- Sample mean is

$$\hat{\mu} = \frac{1}{n}\sum X_i$$

- Keep a record of the sum as the experiment progresses
- Update the sample mean, when needed

# Estimating the mean of a rand. var.

- Sample variance is

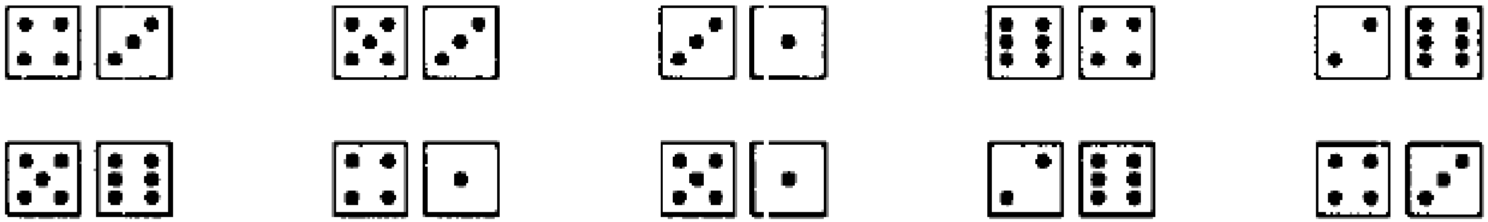
$$\hat{\sigma}^2 = \frac{1}{n-1}\sum X_i^2 - \frac{1}{n(n-1)}\left(\sum X_i\right)^2$$

- Keep a record of the sums as the experiment progresses
- Update the sample variance, when needed

- Estimate the mean as

$$\hat{\mu} \pm \hat{\sigma}/\sqrt{n}$$

# Example – 10 rolls of two dice



- Sample mean of the spot sum

$$\hat{\mu} = 7.4$$

- Sample variance

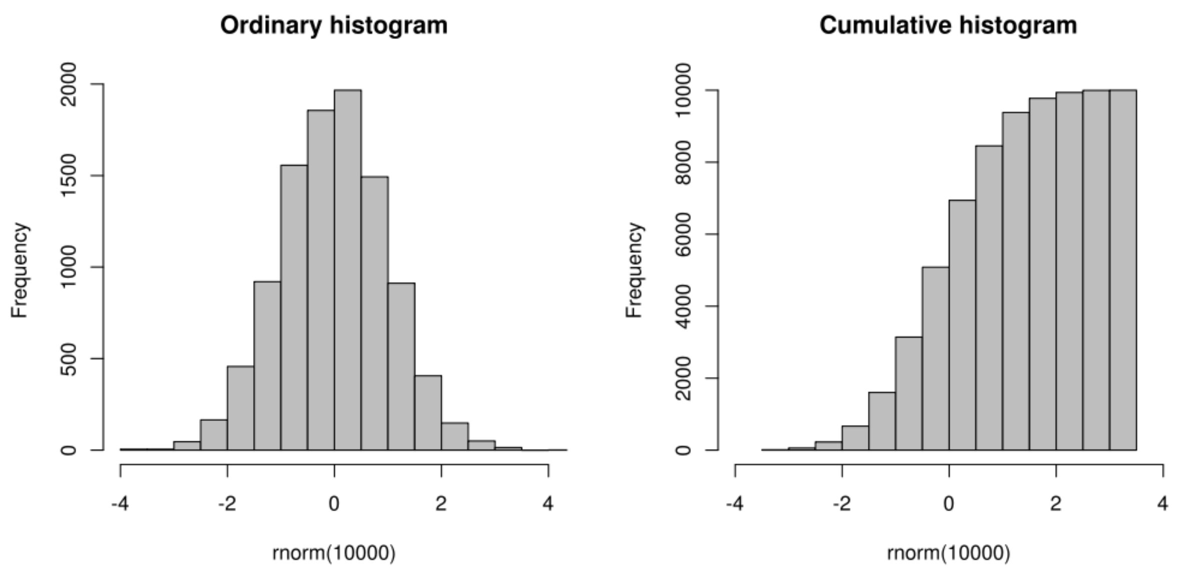
$$\hat{\sigma}^2 \approx 2.1^2$$

- Estimate

$$7.4 \pm 0.7$$

# Histogram

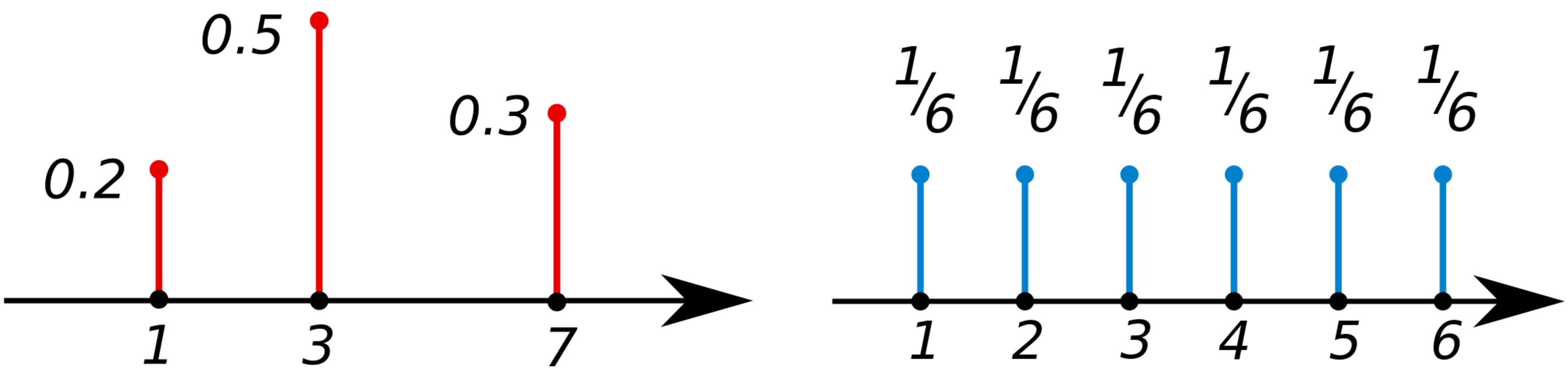
- Graphical representation of the distribution of numerical data
- May be normalized to display relative “frequencies”
- A cumulative histogram represents the cumulative number of observations




[Wikipedia]

# Probability mass function

- Describes the relative likelihood of a discrete random variable to take on a given value




[Wikipedia]

# APPLICATION PROBLEMS

# Task – Problem 1

- Consider the statistical experiment in which a fair coin is tossed repeatedly until one of the faces appears for the second time
- An outcome of the experiment is a list of the faces that appear
  - Possible outcomes: (H, T, H) ; (T, T) ; (H, H) ; (T, H, T)
- Let *Y* denote the random variable that tells how many tosses were necessary to produce a repetition of a face
  - For the examples above: 3, 2, 2, 3
- Simulate an observation of Y

# Task – Problem 2

- Consider the statistical experiment in which a fair die is thrown repeatedly until one of the faces appears for the second time
- An outcome of the experiment is a list of the faces that are thrown
  - Possible outcomes: (3, 6, 2, 6) ; (5, 5) ; (1, 4, 3, 6, 2, 4)
- Let $Y$ denote the random variable that tells how many throws were necessary to produce a repetition of a face
  - For the examples above: 4, 2, 6
- Simulate an observation of Y

# Task – Problem 3

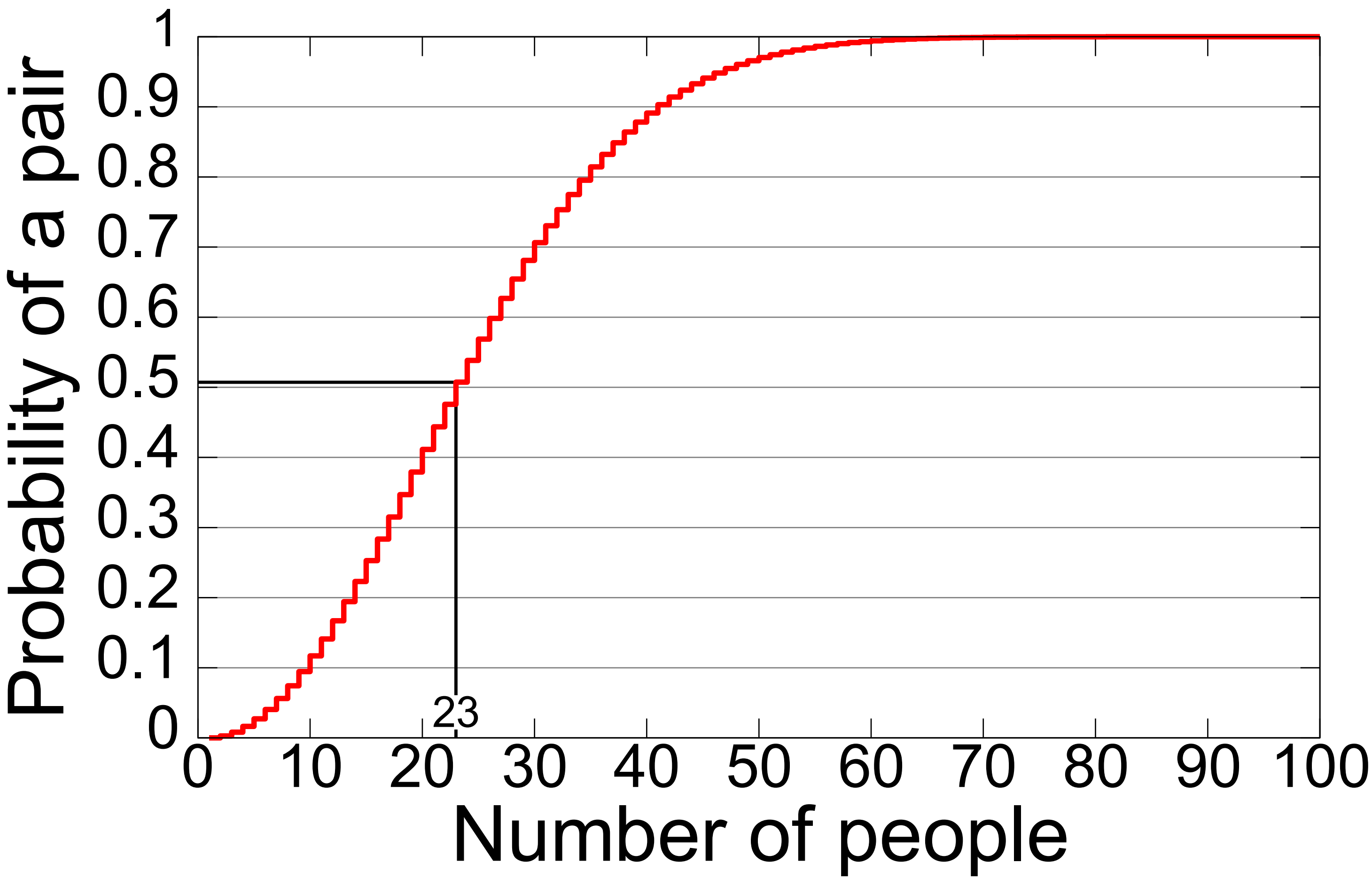
- In a party with $n$ people, what is the probability of at least two of them celebrating their birthday in the same day ?

- What is the smallest $n$ that guaranties that the previous probability is above 50% ?

- Estimate the values with simulated experiments
- Consider that each birthday is equally likely

- Read about "The Birthday Paradox" !

# The Birthday Paradox




[Wikipedia]

# Task – Problem 4

- Consider $n = 4000$
- Generate random numbers in the domain $[n]$ until two have the same value
- How many random trials did that take ?
  - Use $k$ to represent this value
- Repeat the experiment $m = 300$ times, and record for each how many random trials that took

# Task – Problem 4

- Plot that data as a *cumulative density plot*
  - *T*he *x*-axis records the number *k* of trials required, and the *y*-axis records the fraction of experiments that succeeded (a collision) after *k* trials

- Empirically estimate the expected number of *k* random trials in order to have a collision
  - That is, add up all values *k*, and divide by *m*

- How long did it take ?
- Carry out some tests for much larger n and m values !!

# Task – Problem 5

- Consider the statistical experiment in which a fair coin is tossed repeatedly until each face has appeared at least once
- An outcome of the experiment is a list of the faces that appear
  - Possible outcomes: (H, T) ; (T, T, H) ; (H, H, H, T)
- Let *Y* denote the random variable that tells how many tosses were necessary
  - For the examples above: 2, 3, 4
- Simulate an observation of Y

# Task – Problem 6

- Consider the statistical experiment in which a fair die is thrown repeatedly until each face has appeared at least once
- An outcome of the experiment is a list of the faces that appear
  - Possible outcomes: (1, 2, 4, 5, 3, 6) ; (1, 2, 1, 4, 3, 5, 3, 6)
- Let *Y* denote the random variable that tells how many throws were necessary
  - For the examples above: 6, 8
- Simulate an observation of Y

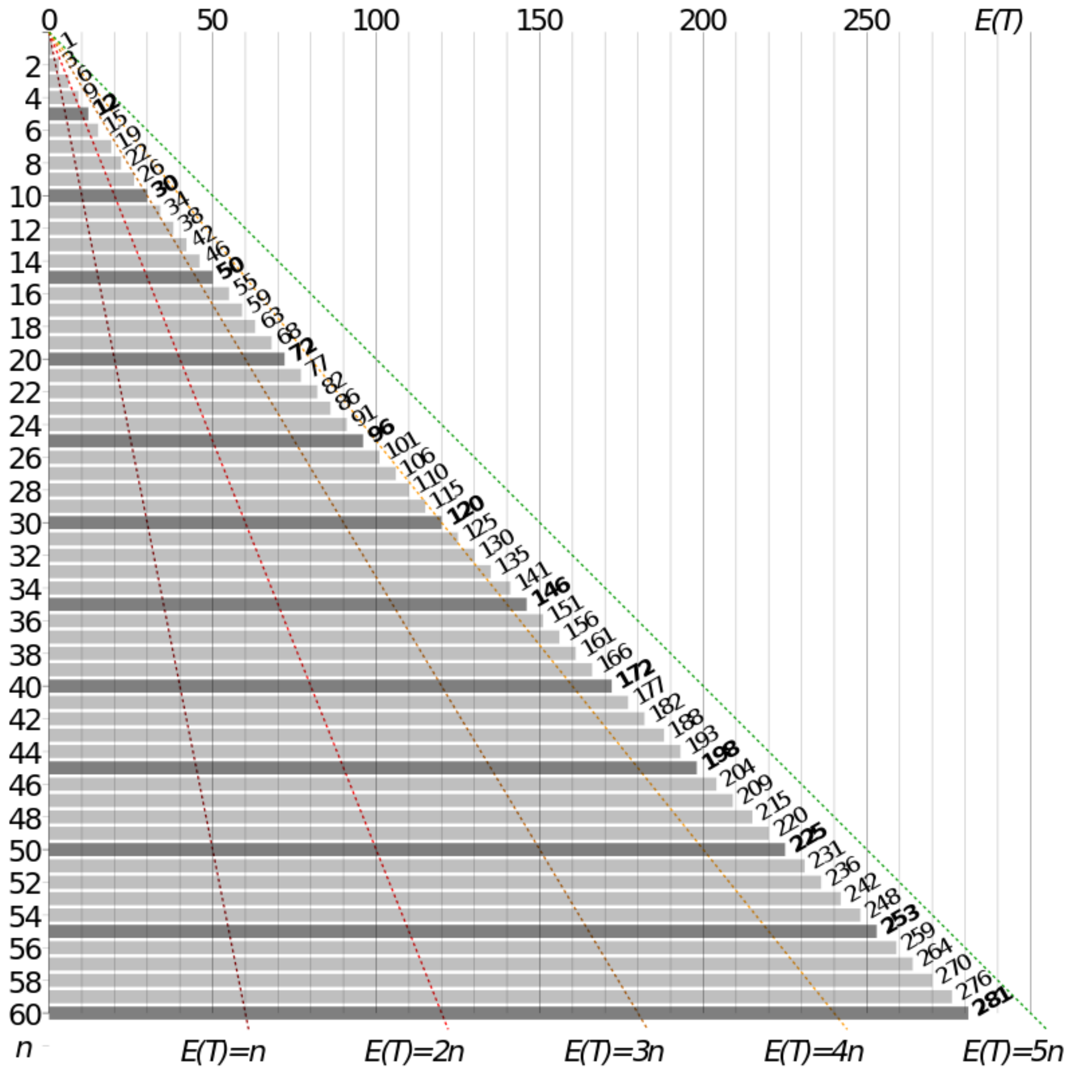
# Task – Problem 7

- Consider *n = 200*
- Generate random numbers in the domain [*n*] until every value *i in* [*n*] has had one random number equal to *i*
- How many random trials did that take ?
  - Use *k* to represent this value
- Repeat the experiment *m* = 300 times, and record for each how many random trials were required to collect all values

# Task – Problem 7

- Plot that data as a *cumulative density plot*
- Empirically estimate the expected number of *k* random trials in order to collect all values
- How long did it take ?
- Carry out some tests for much larger n and m values !!
- Read about "The Coupon Collectors Problem" !!

# The Coupon Collectors Problem




O(n log n)

[Wikipedia]

# Extra Task – Problem 8

- Consider a blind-folded game of darts
- n darts are thrown to m targets
- Each dart reaches one and only one target !

- What is the probability of no target being hit more than once ?

- What is the probability of at least one target being hit at least twice ?

# REFERENCES

# References


- D. Vrajitoru and W. Knight, *Practical Analysis of Algorithms*, Springer, 2014
  - Chapter 6
- R. L. Graham, D. E. Knuth and O. Patashnik, *Concrete Mathematics*, Addison-Wesley, 1989
  - Chapter 8
- J. Hromkovic, *Design and Analysis of Randomized Algorithms*, Springer, 2005
  - Chapter 2

# Acknowledgments

- An earlier version of some of these slides was developed by Professor Carlos Bastos
- Some of the initial examples are from the "Concrete Mathematics" book
- Problems 3 and 4 are from Jeff M. Phillips' Data Mining course at the University of Utah