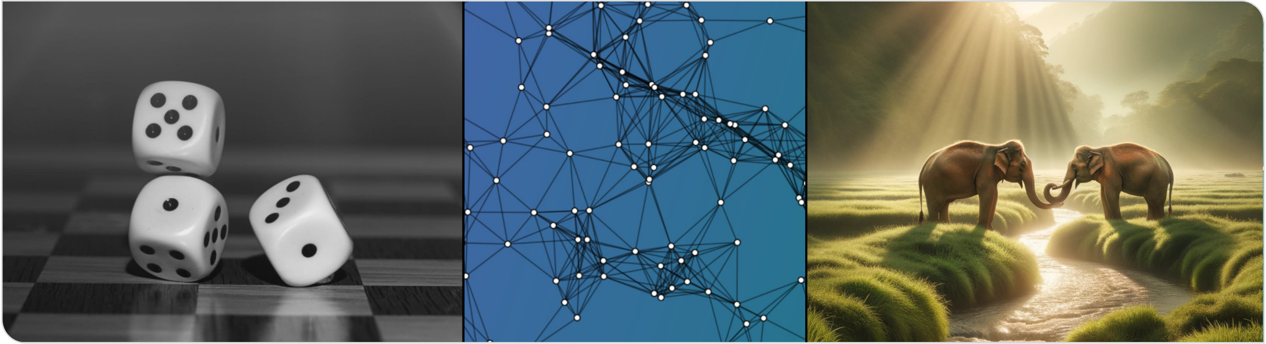
Karlsruhe Institute of Technology

Probability and Computing

Coupling, Balls into Bins, Poissonisation and the Poisson Point Process

Stefan Walzer | WS 2024/2025



Content

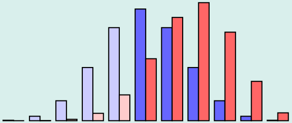
An easy choice?

A Simple Game

You win if you get ≥ 5 heads in 10 coin tosses. Choose:

- i a fair coin with $\Pr[\text{“heads”}] = \frac{1}{2}$
- ii a biased coin with $\Pr[\text{“heads”}] = \frac{2}{3}$

How to prove that (ii) is the better choice?



fair coin

biased coin

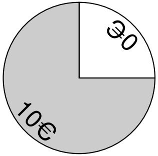
$$\sum_{i=5}^{10} \binom{10}{i} \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{10-i} \stackrel{?}{<} \sum_{i=5}^{10} \binom{10}{i} \left(\frac{2}{3}\right)^i \left(\frac{1}{3}\right)^{10-i}$$

Shouldn’t there be an answer that needs no calculation?

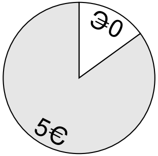
Which Lottery do you prefer?

Karlsruhe Institute of Technology

Consider two “wheels of fortune”:

$X \sim$ 

and

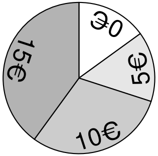
$Y \sim$ 

Both can be rationally preferred

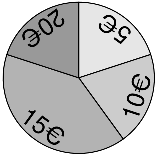
- $\mathbb{E}[X] > \mathbb{E}[Y]$ // maximises expected reward
- $\Pr[Y \geq 5€] > \Pr[X \geq 5€]$ // maximises probability that you can afford ice cream

See `https://en.wikipedia.org/wiki/Von_Neumann%E2%80%93Morgenstern_utility_theorem` to get started on rational choice theory.

Which Lottery do you prefer?

$X \sim$ 

and

$Y \sim$ 

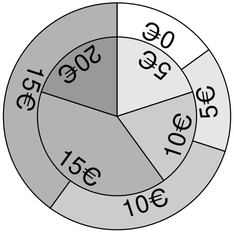
Formal Reason you should prefer Y

For every c we have:

$$\Pr[X \geq c] \leq \Pr[Y \geq c].$$

Intuitive Reason you should prefer Y

Glueing the wheels together guarantees $X < Y$.

$(X, Y) \sim$ 

Content

Equality in Distribution

Notation

We write $X \overset{\mathrm{d}}{=} X'$ for two random variables if X and X' have the same distribution.

Equivalent Definitions

$$X \overset{\mathrm{d}}{=} X' \Leftrightarrow \forall x : \Pr[X = x] = \Pr[X' = x] \quad \text{(for discrete R.V. } X \text{ and } X'\text{)}$$
$$\Leftrightarrow \forall x : \Pr[X \leq x] = \Pr[X' \leq x] \quad \text{(for real-valued R.V. } X \text{ and } X'\text{)}$$

To Clarify:

If $X, Y \sim \mathcal{U}([0,1])$ are independent then

- $X \overset{\mathrm{d}}{=} Y$
- $\Pr[X = Y] = 0$

Definition: Coupling of X and Y

Karlsruhe Institute of Technology

A random variable X $\downarrow$

A random variable Y $\downarrow$

A Coupling of X and Y

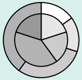
A random variable (X', Y') with

- $X' \overset{d}{=} X$
- $Y' \overset{d}{=} Y$

$X \sim$  $\downarrow$

$Y \sim$  $\downarrow$

A Coupling of X and Y

$(X', Y') \sim$ 

- $X' \overset{d}{=} X$ ✓
- $Y' \overset{d}{=} Y$ ✓

Remarks

- No assumption on joint distribution of X and Y. Might be independent, correlated or undefined.
- X' and Y' should be correlated in an interesting/useful way.
- Example coupling shows:

$$\Pr[X \geq c] \overset{X \overset{d}{=} X'}{=} \Pr[X' \geq c] \overset{X' \leq Y'}{\leq} \Pr[Y' \geq c] \overset{Y \overset{d}{=} Y'}{=} \Pr[Y \geq c]$$

An easy choice!

Karlsruhe Institute of Technology

A Simple Game (Generalised)

You win if your random variable exceeds $c \in \mathbb{N}$. Choose:

- i $X \sim \text{Bin}(n, \frac{1}{2})$ // number of heads of fair coin
- ii $Y \sim \text{Bin}(n, \frac{2}{3})$ // number of heads of biased coin

Prove that Y is better than X using a Coupling

Let $R_1, \ldots, R_n \sim \mathcal{U}([6])$ be n fair dice rolls.

- $X' = |\{i \in [n] \mid R_i \in \{1,2,3\}\}|$
- $Y' = |\{i \in [n] \mid R_i \in \{1,2,3,4\}\}|$

Observe:

- $X' \overset{\text{d}}{=} X$
- $Y' \overset{\text{d}}{=} Y$
- $X' \leq Y'$ guaranteed

Hence: $\Pr[X \geq c] = \Pr[X' \geq c] \leq \Pr[Y' \geq c] = \Pr[Y \geq c]$.

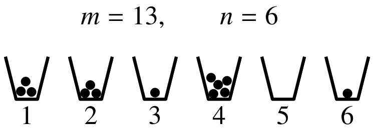
Content

Balls Into Bins

Karlsruhe Institute of Technology

General Terminology

- *m* balls are randomly distributed among *n* bins
- the *load* of a bin is the number of balls in it

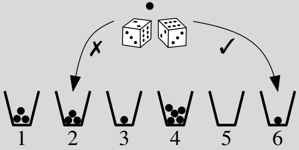


Fully Random Allocation

- $X_1, \ldots, X_m \sim \mathcal{U}([n])$ independent
- $L_i := |\{j \in [m] \mid X_j = i\}|$ is the load of bin $i \in [m]$
- $(L_1, \ldots, L_n)$ follows a (specific) *multinomial distribution*

Example for Partially Random Allocation (not in this lecture)

- balls are placed sequentially
- each ball chooses the *least loaded* among two randomly chosen bins (ties broken randomly)



Balls into Bins: Many Interesting Questions

- What is the expected/distribution/concentration of
  - the load $L_{\max}$ of the most loaded bin
  - the load $L_{\min}$ of the least loaded bin
  - $L_{\max} - L_{\min}$
  - the number of empty bins
  - ...
- Can we make the allocation more balanced by intervening in some way?
  - e.g. with partially random allocation from last slide $\mathbb{E}[L_{\max} - L_{\min}]$ stays bounded when $m \to \infty$ while n is fixed.

Countless variants exist...

“Balls into Bins” is everywhere

Karlsruhe Institute of Technology

Hashing with Chaining $\longleftrightarrow$ n Balls into m Bins

length of the list in bucket i $\longleftrightarrow$ number of balls in bin i

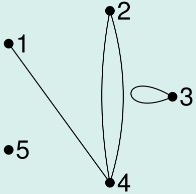
Bloom Filter with k Hash Functions $\longleftrightarrow$ kn Balls into m Bins

a filter bit is set to 1 $\longleftrightarrow$ ith bin is non-empty

Degree Sequence of Random (Multi-)Graph $\longleftrightarrow$ $2m$ Balls into n bins

Given independent $v_1, \ldots, v_{2m} \sim \mathcal{U}([n])$ let
$G = (V = [n], E = \{\{v_1, v_2\}, \ldots, \{v_{2m-1}, v_{2m}\}\})$
(we allow multiedges and loops in G)

degree of vertex i $\longleftrightarrow$ load of bin i



$n = 5, \quad m = 4$

$v_1 = 1, \quad v_2 = 4$
$v_3 = 4, \quad v_4 = 2$
$v_5 = 3, \quad v_6 = 3$
$v_7 = 2, \quad v_8 = 4$

“Balls into Bins” is the standard language for discussing underlying mathematical questions.

Content

Load of a Single Bin

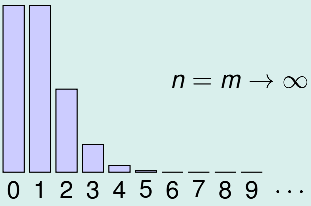
Karlsruhe Institute of Technology

Setting: Expected Constant Load per Bin

- fully random allocation
- $m = \lambda n$ balls n bins for large n
- λ fixed constant

Load of the First Bin

Consider $L^{(n)} \sim \text{Bin}(\lambda n, \frac{1}{n})$. For $\lambda = 1$:



Poisson Distribution

For $\lambda \in \mathbb{R}_{\geq 0}$, $X \sim \text{Pois}(\lambda)$ is a random variable with

$$\Pr[X = i] = e^{-\lambda} \frac{\lambda^i}{i!}$$ // note: probabilities sum to 1

Theorem (proof on blackboard)

$\lim_{n \to \infty} \Pr[L^{(n)} = i] = \Pr[X = i]$.

Remarks

- we say "$L^{(n)}$ converges in distribution to X"
- we write $L^{(n)} \xrightarrow{d} X$
- this formally refers to convergence of CDFs

Properties of the Poisson Distribution

Karlsruhe Institute of Technology

Exercise: $X \sim \mathrm{Pois}(\lambda)$ has Nice Properties

- i $\mathbb{E}[X] = \lambda$.
- ii $\mathrm{Var}(X) = \lambda$.
- iii Let $Y \sim \mathrm{Pois}(\rho)$ be independent of X. Then $X + Y \sim \mathrm{Pois}(\lambda + \rho)$.
- iv Let $X' \sim \mathrm{Bin}(X, p)$. Then $X' \sim \mathrm{Pois}(\lambda p)$.

Poissonised Balls into Bins

$(L_i):\ 3\quad 3\quad 1\quad 5\quad 0\quad 1$

λn Balls into n Bins Model

- $X_1, \ldots, X_{\lambda n} \sim \mathcal{U}([n])$
- $L_i := |\{j \in [m] \mid X_j = i\}| \sim \mathrm{Bin}(\lambda n, \frac{1}{n})$
- $(L_i)_{i \in [n]}$ *not independent*
  - e.g. large L_1 is (weak) evidence for small L_2
  - annoying in analysis
- number λn of balls *fixed*

"Poissonised" Model

- $L_1, \ldots, L_n \sim \mathrm{Pois}(\lambda)$ independent
  - extremely convenient for analysis
- $\mathbb{E}[L_1 + \cdots + L_n] = \lambda n$
- number of balls *random* $\sim \mathrm{Pois}(\lambda n)$
  - unusual setting in practice

Wouldn't it be nice…

… if we could switch between the models whenever convenient?

Connection: Poissonised and Regular Balls into Bins

Karlsruhe Institute of Technology

Lemma 1

Let $n \in \mathbb{N}$ and $\lambda > 0$. Consider two variants of Poissonised balls into bins:

Regular Variant:

- sample $L_1, \ldots, L_n \sim \text{Pois}(\lambda)$

Sum-First-Variant:

- sample $M \sim \text{Pois}(\lambda n)$
- perform a regular M balls into n bins experiment
  - sample $X_1, \ldots, X_M \sim \mathcal{U}([n])$
  - let $L'_i := |\{j \in [M] \mid X_j = i\}|$

Both variants are equivalent, i.e. $(L_1, \ldots, L_n) \overset{d}{=} (L'_1, \ldots, L'_n)$.

What we need to show (calculation on blackboard):

For arbitrary $(\ell_1, \ldots, \ell_n) \in \mathbb{N}^n$: $\Pr[(L_1, \ldots, L_n) = (\ell_1, \ldots, \ell_n)] = \Pr[(L'_1, \ldots, L'_n) = (\ell_1, \ldots, \ell_n)]$.

Some Concentration Bounds

Karlsruhe Institute of Technology

Lemma 2

- i Let $\Lambda > 0$ and $X \sim \text{Pois}(\Lambda)$. Then $\Pr[|X - \Lambda| > t] \leq \frac{\Lambda}{t^2}$ for any $t > 0$. // Chebyschev
- ii Let $\lambda = \Theta(1)$, and $X \quad \sim \text{Pois}(\lambda \quad n)$ then $\Pr[X \quad = \lambda n \pm \mathcal{O}(n^{2/3})] = 1 - o(1)$.
- iii Let $\lambda = \Theta(1)$, $\lambda^+ := \lambda + n^{-1/3}$ and $X^+ \sim \text{Pois}(\lambda^+ n)$ then $\Pr[X^+ \geq \lambda n] = 1 - o(1)$.
- iv Let $\lambda = \Theta(1)$, $\lambda^- := \lambda - n^{-1/3}$ and $X^- \sim \text{Pois}(\lambda^- n)$ then $\Pr[X^- \leq \lambda n] = 1 - o(1)$.
- v In particular: $\Pr[X^- \leq \lambda n \leq X^+] = 1 - o(1)$.

Coupling of Poissonised and Regular Balls into Bins

Karlsruhe Institute of Technology

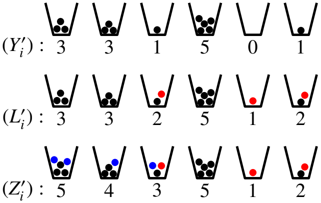
Theorem

Let $n, \lambda, \lambda^+, \lambda^-$ be as before. Consider three "balls into bins" models:

1. $Y_1, \dots, Y_n \sim \text{Pois}(\lambda^-)$ // poissonised with reduced λ
2. $L_1, \dots, L_n$ arising from regular $m = \lambda n$ balls into n bins
3. $Z_1, \dots, Z_n \sim \text{Pois}(\lambda^+)$ // poissonised with increased λ

There is a coupling $(Y_i', L_i', Z_i')_{i\in[n]}$ of $(Y_i)_{i\in[n]}, (L_i)_{i\in[n]}, (Z_i)_{i\in[n]}$ such that

$$\text{with probability } 1 - o(1)\text{: } Y_i' \le L_i' \le Z_i' \text{ for all } i \in [n].$$



Proof.

Let $X_1, X_2, \dots \sim \mathcal{U}([n])$, $M^- \sim \text{Pois}(\lambda^- n)$, $M^+ \sim \text{Pois}(\lambda^+ n)$.

- $Y_i' := |\{j \in [M^-] \mid X_j = i\}|$ for $i \in [n]$.
- $L_i' := |\{j \in [m\ \ ] \mid X_j = i\}|$ for $i \in [n]$.
- $Z_i' := |\{j \in [M^+] \mid X_j = i\}|$ for $i \in [n]$.

This is indeed a coupling as claimed:

- $(Y_i')_{i\in[n]} \overset{d}{=} (Y_i)_{i\in[n]}$ by Lemma 1.
- $(L_i')_{i\in[n]} \overset{d}{=} (L_i)_{i\in[n]}$ by construction.
- $(Z_i')_{i\in[n]} \overset{d}{=} (Z_i)_{i\in[n]}$ by Lemma 1.

By the Corollary we have $M^- \le m \le M^+$ with probability $1 - o(1)$. In that case clearly $Y_i' \le L_i' \le Z_i'$ for all $i \in [n]$. $\square$

Coupling of Poissonised and Regular Balls into Bins

Karlsruhe Institute of Technology

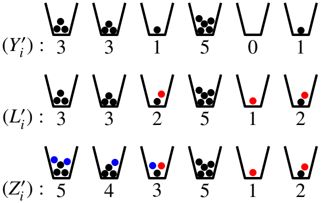
Theorem

Let $n, \lambda, \lambda^+, \lambda^-$ be as before. Consider three "balls into bins" models:

1. $Y_1, \ldots, Y_n \sim \mathrm{Pois}(\lambda^-)$ // poissonised with reduced λ
2. $L_1, \ldots, L_n$ arising from regular $m = \lambda n$ balls into n bins
3. $Z_1, \ldots, Z_n \sim \mathrm{Pois}(\lambda^+)$ // poissonised with increased λ

There is a coupling $(Y'_i, L'_i, Z'_i)_{i \in [n]}$ of $(Y_i)_{i\in[n]}, (L_i)_{i\in[n]}, (Z_i)_{i\in[n]}$ such that

$$\text{with probability } 1 - o(1)\text{: } Y'_i \le L'_i \le Z'_i \text{ for all } i \in [n].$$



Application involving Monotonous Functions

Let $f : \mathbb{N}_0^n \to \mathbb{R}$ be non-decreasing in each argument.
Examples:

- maximum load of a bin
- longest run of non-empty bins
- collision number // numbers of pairs of co-located balls

For some bound $B \in \mathbb{R}$ let

- $p^- := \Pr[f((Y_i)_{i\in[n]}) \ge B]$ // easier to compute
- $p := \Pr[f((L_i)_{i\in[n]}) \ge B]$ // what we want
- $p^+ := \Pr[f((Z_i)_{i\in[n]}) \ge B]$ // easier to compute

Then $p \in [p^- - o(1), p^+ + o(1)]$.

Back to Bloom Filters

Exercise:

Analyse Bloom filters in a "Poissonised" model and discuss how the results can be transferred to the exact model.

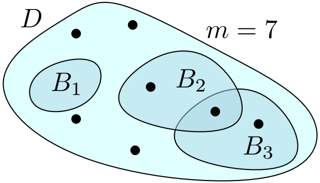
Content

What about "Balls into Continuous Domain"?

Karlsruhe Institute of Technology

Setting

- D is space of finite measure
- $m \in \mathbb{N}$ // number of balls
- $X_1, \ldots, X_m \sim \mathcal{U}(D)$ // randomly thrown into D



Note: If $D = \{1, \ldots, n\}$ we have discrete balls into bins.

Same annoying issue

If $B_1, B_2 \subseteq D$ with $B_1 \cap B_2 = \varnothing$ are two "bins" then the numbers L_1 and L_2 of "balls" in B_1 and B_2 are correlated.

Similar elegant solution

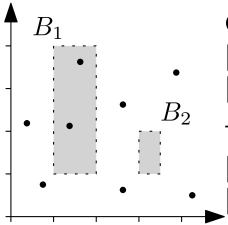
- We can "Poissonise" the setting.
- But we drop "balls into bins" terminology:
  - we allow infinite domains D
  - we allow infinite number of balls

Definitions of the Poisson Point Process

General Definition

Let D be a measurable space with measure μ. // e.g. $D = \mathbb{R}^2$ and $\mu =$"area"
The Poisson point process with parameter $\lambda \in \mathbb{R}_{\geq 0}$ is a random set $P \subseteq D$ such that

1. $|P \cap B| \sim \text{Pois}(\lambda\mu(B))$ for any $B \subseteq D$ with $\mu(B) < \infty$
2. $|P \cap B_1|$ and $|P \cap B_2|$ are independent whenever $B_1 \cap B_2 = \varnothing$



Generally:
$|B_1 \cap P| \sim \text{Po}(3\lambda)$
$|B_2 \cap P| \sim \text{Po}(\frac{1}{2}\lambda)$

This outcome:
$|B_1 \cap P| = 2$
$|B_2 \cap P| = 0$

Equivalent Definition if $\mu(D) < \infty$

- sample $M \sim \text{Pois}(\lambda\mu(D))$
- sample $X_1, \ldots, X_M \sim \mathcal{U}(D)$

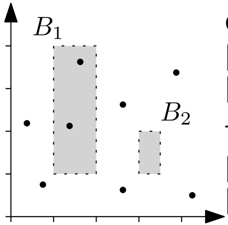
Then $P \overset{d}{=} \{X_1, X_2, \ldots, X_M\}$.

Definitions of the Poisson Point Process

General Definition

Let D be a measurable space with measure μ. // e.g. $D = \mathbb{R}^2$ and $\mu =$"area"
The Poisson point process with parameter $\lambda \in \mathbb{R}_{\geq 0}$ is a random set $P \subseteq D$ such that

1. $|P \cap B| \sim \text{Pois}(\lambda\mu(B))$ for any $B \subseteq D$ with $\mu(B) < \infty$
2. $|P \cap B_1|$ and $|P \cap B_2|$ are independent whenever $B_1 \cap B_2 = \varnothing$

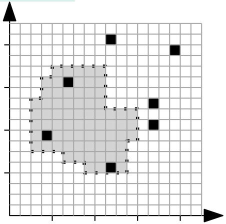


Generally:
$|B_1 \cap P| \sim \text{Po}(3\lambda)$
$|B_2 \cap P| \sim \text{Po}(\frac{1}{2}\lambda)$

This outcome:
$|B_1 \cap P| = 2$
$|B_2 \cap P| = 0$

Construction as a limit

- subdivide D into pieces of measure ε
- let each piece contain a point with probability $\varepsilon\lambda$
- consider the limit for $\varepsilon \to 0$



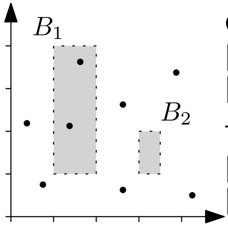
$$|B \cap P| \sim \text{Bin}(\mu(B)/\epsilon, \lambda\epsilon) \xrightarrow{\epsilon \to 0} \text{Po}(\lambda\mu(B))$$

Definitions of the Poisson Point Process

General Definition

Let D be a measurable space with measure μ. // e.g. $D = \mathbb{R}^2$ and $\mu =$"area"
The Poisson point process with parameter $\lambda \in \mathbb{R}_{\geq 0}$ is a random set $P \subseteq D$ such that

1. $|P \cap B| \sim \text{Pois}(\lambda\mu(B))$ for any $B \subseteq D$ with $\mu(B) < \infty$
2. $|P \cap B_1|$ and $|P \cap B_2|$ are independent whenever $B_1 \cap B_2 = \varnothing$

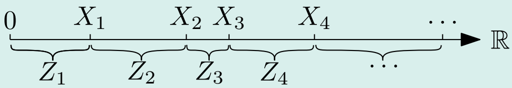


Generally:
$|B_1 \cap P| \sim \text{Po}(3\lambda)$
$|B_2 \cap P| \sim \text{Po}(\frac{1}{2}\lambda)$

This outcome:
$|B_1 \cap P| = 2$
$|B_2 \cap P| = 0$

Equivalent Definition if $D = \mathbb{R}_{\geq 0}$ (where μ is the Borel measure)

- sample $Z_1, Z_2, \ldots \sim \text{Exp}(\lambda)$
- define $X_i = \sum_{j=1}^{i} Z_j$



Then $P \stackrel{d}{=} \{X_1, X_2, \ldots\}$.

Proof idea: $\Pr[\min P > t] = \Pr[|P \cap [0, t]| = 0] = \Pr_{X \sim \text{Pois}(\lambda t)}[X = 0] = e^{-\lambda t} \stackrel{\text{def}}{=} \Pr[Z_1 > t]$.

Conclusion

Karlsruhe Institute of Technology

Coupling

- embedding of two random variables X and Y into a common probability space
- relationships between distributions of X and Y become visible as relationships between outcomes of X' and Y'

Balls into Bins

- standard language when m objects are randomly assigned to n other objects

Poissonisation

- the act of replacing multinomially distributed $(L_1, \ldots, L_n)$ with independent Poisson random variables $(L'_1, \ldots, L'_n)$
- often results in model with nicer mathematical properties
- often formally justifiable

Poisson Point Process

- important model where points from a continuous space occur independently from each other with fixed density λ

Anhang: Mögliche Prüfungsfragen I

- Was ist ein Coupling?
  - Nenne Beispiele in denen ein Coupling nützlich sein kann.
  - Was bedeutet Gleichheit in Verteilung?
- Wo in der Vorlesung haben wir (implizit oder explizit) Balls-into-Bins-Prozesse betrachtet?
- Poissonisierung:
  - Welche lästige Eigenschaft hat die Verteilung der Beladungen in Balls-into-Bins Prozessen? Was ist in einem poissonisierten Modell anders?
  - Wie lässt sich in einem Balls-into-Bins Setting die Poissonverteilung wiederfinden?
  - Wie haben wir das poissonisierte und das reguläre Balls-into-Bins-Modell miteinander in Verbindung gebracht? Inwiefern lässt sich ein Wechsel zwischen den Modellen formal rechtfertigen?
- Poisson-Punktprozesse
  - Wie sind Poisson-Punktprozesse definiert?