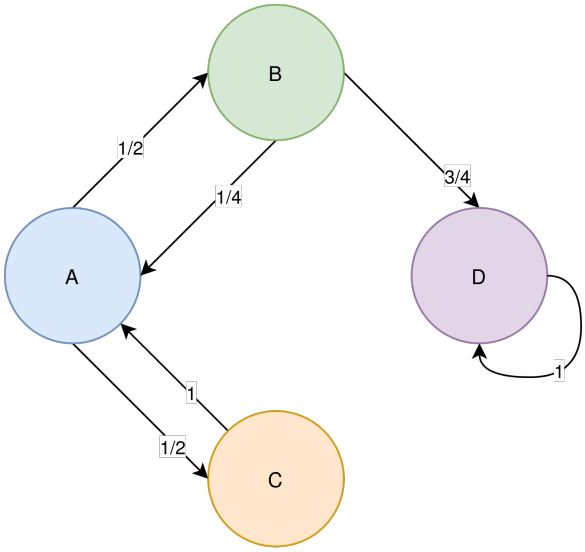
Probability and statistics 2

Filip Mihál', Tomáš Turek



October 15, 2023

INFO

This is shortened text from Probability and statistics 2. Almost all examples are omitted since this is just shortened version. Also you may find several notes of what was omitted and what is the reasoning behind that. Usually it is the lack of my time.

Contents

1. Markov chains

"Some type of an automata, that represent probability space. It needs to have special properties."

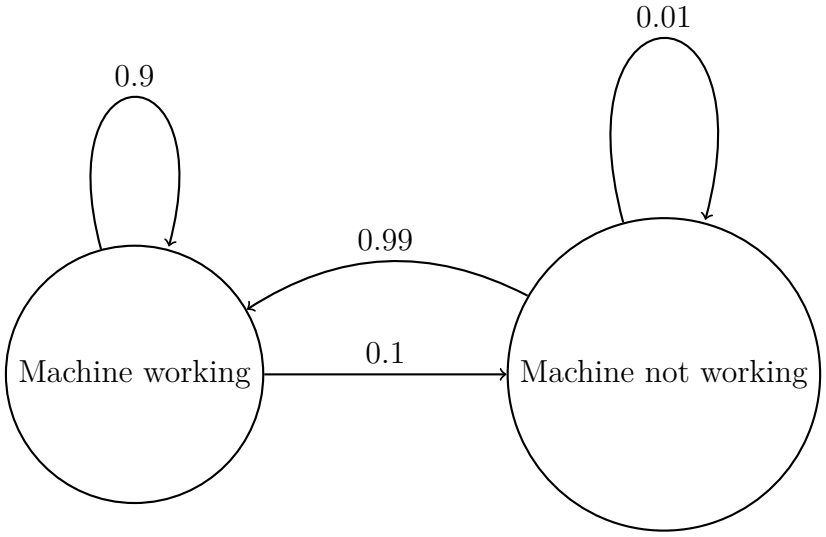


Figure 1.1: Example of a Markov chain.

1.1 Model

- States: S it is usually finite and sometimes only countable.
- sequence $X_0, X_1, X_2, \ldots$ of random variables with values in S
- X_{t+1} depends **only** on X_t
- $\Pr[X_{t+1} = j | X_t = i] = p_{ij}$ where $i,j \in S$

Definition 1. *Sequence of r.v.* X_0,X_1,$X_2, \ldots$ *is a* ***Markov chain*** *if:*

- $\exists$ *countable* S : *Rng* $X_t \subset S$ $\forall t$
- $\forall t \in \mathbb{N}$ $\forall a_0, a_1, a_2, \ldots, a_{t+1} \in S$

$$\Pr[X_{t+1} = a_{t+1} | X_0 = a_0, X_1 = a_1, \ldots, X_t = a_t] = \Pr[X_{t+1} = a_{t+1} | X_t = a_t]$$

This means that Markov chain has the property of being memory-less and this probability written above is called *transition probability*. We can map all elements from S to a number from range 1,2,...,n and then we can build **transition matrix**.

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & & \\ p_{31} & & \ddots & \\ \vdots & & & \end{pmatrix}$$

Where p_{ij} means going from i to j. All $p_{ij} \geq 0$ and the sum of each row is 1. Also we can build **transition graph** representing this Markov chain. In that graph $V = S$ and arcs exists if $(ij) : p_{ij} > 0$.

Now we look at distribution, or PMF of $X_k = \pi^{(k)}$ where $\pi^{(k)} = \left(\pi_1^{(k)}, \pi_2^{(k)}, \pi_3^{(k)}, \dots\right)$ and the sum is 1. Then we may see that $\pi_j^{(k)} = \Pr[X_k = j]$. We will be calling $\pi^{(0)}$ an *initial state*.

Then we can see that $\pi^{(1)} = \pi^{(0)}P$ as multiplication by transition matrix. We can generalize this to:

$$\pi^{(k)} = \pi^{(k-1)}P$$

Theorem 1. *For any MC with transition matrix P we have $\pi^{(k)} = \pi^{(0)}P^k$ and $\pi^{(k+1)} = \pi^{(t)}P^k$.*

Proof. Proof will be by induction. So $\pi^{(k+1)} = \pi^{(k)}P = \pi^{(0)}P^kP = \pi^{(0)}P^{k+1}$. □

Definition 2. ***K-step transition*** *is defined as:*

$$\begin{array}{rcl} r_{ij}(k) & := & \Pr[\textit{from } i \textit{ to } j \textit{ in } k \textit{ steps}] \\ & = & \Pr[X_k = j | X_0 = i] \\ & = & \Pr[X_{t+k} = j | X_t = i] \\ r_{ij}(1) & = & p_{ij} \end{array}$$

Observation.

$$r_{ij}(k) = \pi_j^{(k)} \textit{ if } \pi^0 = (0, 0, \dots, 0, 1, 0 \dots, 0)$$

Where 1 *is on i-th position. Also:*

$$\pi_j^{(k)} = (\pi^0 P^k)_j = \left((0, 0, \dots, 0, 1, 0 \dots, 0)\, P^k\right)_j = (P^k)_{ij}$$

1.2 Chapman-Kologorov formula

$$\begin{array}{rcl} r_{ij}(k) & = & (P^k)_{ij} \\ r_{ij}(k+l) & = & \sum_{t=1}^{S} r_{it}(k) r_{tj}(l) \\ r_{ij}(k+1) & = & \sum_{t=1}^{S} r_{it}(k) p_{tj} \end{array}$$

Definition 3. *j is* ***accessible*** *from i if*

$$\begin{array}{c} (j \in A(i), i \to j) \\ \Updownarrow \\ \Pr[\exists k \geq 0 : X_k = j | X_0 = i] > 0 \\ \Updownarrow \\ \sum_{k=0}^{\infty} r_{ij}(k) > 0 \\ \Updownarrow \\ \exists \textit{ a discrete path from } i \textit{ to } j \\ \textit{in the transition graph} \end{array}$$

Definition 4. *i and j from S are* ***commuting states*** *$(i \leftrightarrow j)$ iff $i \to j$ and $j \to i$.*

Lemma 1. *$\leftrightarrow$ is an equivalence relation.*

Proof. We need to show that it satisfies reflexivity, symmetry and transitivity.

1. $i \leftrightarrow i$ which means $i \to i$ so $r_{ii}(0) = 1$
2. $i \leftrightarrow j$ iff $j \leftrightarrow i$ by definition
3. $i \leftrightarrow j$ and $j \leftrightarrow t$ we want to show $i \leftrightarrow t$, but we know $i \to j \to t$ and $t \to j \to i$ so we use these paths (or just shorten them by first intersection).

□

Definition 5. *An **equivalence class** in a Markov chain is a set of states that are commuting with each other. The set is maximal with its property. In other words, no additional state from S can be included in the set without breaking the commuting property.*

Definition 6. *MC is called **irreducible** if $\leftrightarrow$ has just 1 equivalence class. This is equivalent to that $\forall ij : i \leftrightarrow j$.*

Or by graph theory we can say that the transition graph is strongly connected and when we compress these classes we get DAG.

Definition 7. *$i \in S$ is called **recurrent** if $\forall j \in A(i) : i \in A(j)$ and **transient** otherwise.*

Theorem 2. *$i \in S$ we define $f_{ii} = \Pr[\exists t \geq 1 : X_t = i | X_0 = i]$ or by words "probability of going back to i". Then:*

- *i is recurrent iff $f_{ii} = 1$*
- *i is transient iff $f_{ii} < 1$*

Proof. i is transient iff $\exists j \in A(i) : i \notin A(j)$. Starting with $X_0 = i$ the probability $\exists t \geq 1 : X_t = j$ is $p > 0$ and $\Pr[\text{going to } i \text{ from } j] = 0 \Rightarrow f_{ii} \leq 1 - p$. And if i is recurrent then $f_{ii} = 1$. □

Definition 8. *$i \in S$ we define V_i as number of visits to i or written as $|\{t : X_t = i\}|$*
$V_i \in \mathbb{N} \cup \{\infty\}$ so it is a random variable defined by $X_0, X_1, \ldots$

Theorem 3. *i is recurrent $\Rightarrow \Pr[V_i = \infty | X_0 = i] = 1$*
i is transient $\Rightarrow (V_i | X_0 = i) \sim Geom(1 - f_{ii})$, where $(1 - f_{ii})$ is called as escape probability.

1.3 Steady state

Definition 9. *Let π be a distribution on S such that $(\pi_1 + \pi_2 + \cdots + \pi_S = 1, \pi_i > 0)$. Then π is **stationary** distribution if $\pi P = \pi$. Or can be written as*
$[\pi = (\pi_1, \pi_2, \ldots) | \forall j \pi_j = \sum_{i \in S} \pi_i p_{ij}]$ for MC with transition matrix P.

Observation. *If $\pi^{(0)} = \pi$ and π is stationary then $\pi^{(1)} = \pi$ and $\forall k : \pi^{(k)} = \pi$.*

Definition 10. *$s \in S$ is **periodic** if $\exists \Delta \geq 2$ integer such that $\Pr[X_t = s | X_0 = s] > 0 \mathbb{R} ightarrow \Delta | t$. MC is periodic if all its states are periodic, otherwise it is aperiodic.*

Theorem 4. *$(X_t)_{t=0}^{\infty}$ is a MC that is irreducible, aperiodic and $|S| < \infty$. Then $\exists \pi$ that is a stationary distribution and $\forall j \forall i \lim_{k \to \infty} r_{ij}(k) = \pi_j$, π is a unique solution to*

$$\pi P = \pi$$

$$\pi \mathbf{1} = 1$$

1.4 Absorption probability

Definition 11. *Absorption states are such states, that the probability of staying in the same state is* 1*. Or it is* $\{s \in S : p_{ss} = 1\}$*.*

Lemma 2 (Probability of Absorption)**.** *Assume a MC with absorbing state* 0 *(and some move). Put*

$$a_i = \Pr[\exists t : X_t = 0 | X_0 = i] \text{ for } i \in S$$

Then (a_i) *are the unique solution to:*

$$\begin{array}{lcll} a_0 & = & 1 & \\ a_i & = & 0 & \text{if } i \neq 0 \text{ and absorbing} \\ a_i & = & \sum_{j \in S} p_{ij} a_j & \text{for } i \text{ not absorbing} \end{array}$$

Proof. $a_0 = 1$ and $a_i = 0$ if $i \neq 0$ and absorbing is easy observation. Lets assume i is not absorbing then

$$\begin{array}{rcl} a_i & = & \Pr[\exists t : X_t = 0 | X_0 = i] = \\ & = & \sum_{j \in S} \Pr[X_1 = j | X_0 = i] \cdot \Pr[\exists t : X_t = 0 | X_0 = i, X_1 = j] = \\ & = & \sum_{j \in S} p_{ij} \Pr[\exists t : X_t = 0 | X_0 = j] = \\ & = & \sum_{j \in S} p_{ij} a_j \end{array}$$

□

1.5 Mean time to absorption

$A \subseteq S$ is set of all absorption states. $T = \min\{t \geq 0 | X_t \in A\}$ is *absorption time* and random variable. Then we define $\mu_i = \mathbb{E}[T | X_0 = i]$.

Theorem 5. $(\mu_i)_{i \in S}$ *is the unique solution to:*

$$\begin{array}{lcl} \text{if } i \in A \quad \mu_i & = & 0 \\ \text{if } i \notin A \quad \mu_i & = & 1 + \sum_{j \in S} p_{ij} \mu_j \end{array}$$

1.6 SAT

Problem where there is given a Boolean formula and we have to say if it is satisfiable.

1.6.1 2-SAT (polynomial)

Special case of SAT where all clauses have at most 2 literals.

Algorithm for 2-SAT

1. Start with any assignment ($x_1 = x_2 = \cdots = x_n = F$)
2. Repeat up to $2mn^2$ times (n is the number of variables and m is an arbitrary parameter)
   - if φ is satisfiable return "YES"

- otherwise, choose any clause that is not satisfied and randomly change one of its variables $(*)$

3. Return "NO"

Pr[incorrectly saying no] $\leq \frac{1}{2m}$ which can be proved by Markov inequality.
Pr[incorrectly saying no] $\leq \frac{1}{2^m}$ using iterative Markov inequality.

1.6.2 3-SAT

Algortihm for 3-SAT

- Repeat for $\leq m$ times
  - Repeat for $\leq 3^{n/2}$ times
    - randomly initialize the variables
    - if φ is satisfiable return "YES"
    - otherwise, choose any clause that is not satisfied and randomly change one of its variables

Running time of this algortihm is exponential in n.
P[failure] $\leq \frac{1}{2^m}$
By using a better algorithm we can get the exponential part to be $\frac{4^n}{3}$.
The idea behind these algorithms is that we are using a *random walk* on the space of all possible assignments. This is a *Markov chain*. So we can easily calculate the probability of getting to the absorbing state and the mean time to get there.

2. Bayesian statistics

2.1 What is probability?

We may look at probability from different angles.

1. Math concepts.
   - axioms, examples $\frac{\#\text{good}}{\#\text{all}}$, theorems ...
   - interesting/useful probabilistic method as "to show $A \neq 0$ we show $\Pr[A] > 0$", lower bounds for Ramsey number
2. Description of real world. Question: *Does Nature play dice?*
   - YES, if quantum theory is right so called *true randomness*
   - imprecise measurements so called *pseudo randomness*

Then we we have two possible approaches.

1. **Frequntist's approach** $\frac{\#\text{ good}}{\#\text{ all}}$
2. **Bayesian approach** as subjective probability, so we are counting with all possible universes and what is the probability this will happen in our universe.

2.2 Bayesian statistics

1. Θ is random variable describing some quantity of interest
2. $X = (X_1, \ldots, X_n)$ measurements

Remark. In Frequentist's approach Θ does not exist we have ϑ as unknown fixed parameter.

2.2.1 Bayes theorem

$$\Pr[B|A] = \frac{\Pr[B]\Pr[A|B]}{\Pr[A]}$$

Where $\Pr[A], \Pr[B] > 0$. We will consider B as $\Theta = \vartheta$ and A as measurements $X = x$. Now we get:

$$\Pr[\Theta = \vartheta | X = x] = \frac{\Pr[\Theta = \vartheta]\Pr[X = x|\Theta = \vartheta]}{\Pr[X = x]}$$

Where $\Pr[\Theta = \vartheta | X = x]$ is called **posterior** and it is the probability after some measurements. $\Pr[\Theta = \vartheta]$ is called **prior** as an probability and $\Pr[X = x|\Theta = \vartheta]$ is our current model of the world (*likelihood*).

variable	PMF	PDF
1	p_Θ	f_Θ
2	p_X	f_X

2.2.2 Bayes theorem using PMF

$$p_{\Theta|X}(\vartheta|x) = \frac{p_\Theta(\vartheta)p_{X|\Theta}(x|\vartheta)}{\sum_{\vartheta'} p_\Theta(\vartheta')p_{X|\Theta}(x|\vartheta')} = cp_\Theta(\vartheta)p_{X|\Theta}(x|\vartheta)$$

For some constant c.

2.3 What do we want?

1. Point estimate for Θ.
2. Interval estimate for Θ.
3. Hypothesis testing.

For *interval estimate* we have given X and want to find $[a,b]$ as $(a = a(X), b = b(X))$. $\Pr[a(x) < \Theta < b(X)|X = x] \geq 1 - \alpha$. Perhaps $\Pr[\Theta < a(x)|X = x] = \frac{\alpha}{2}$ and $\Pr[\Theta > b(x)|X = x] = \frac{\alpha}{2}$.

For *point estimate* we have two approaches.

2.3.1 1) MAP as for maximum aposteriori probability

$$\hat{\vartheta} = \arg\max_{\vartheta} p_{\Theta|X}(\vartheta|x)$$

If $X = x$ what is the most likely value?

2.3.2 2) LMS as for least mean square

$$\begin{aligned} \hat{\vartheta} &= \arg\min_\vartheta \mathbb{E}[(\Theta - \vartheta)^2|X = x] \\ &= \arg\min_\vartheta \mathbb{E}[\Theta|X = x] \end{aligned}$$

2.4 Naive Bayes

By the Bayesian statistics we get for X_1:

$$p_{\Theta|X_1}(\vartheta|x_1) = \frac{p_\Theta(\vartheta)p_{X_1|\Theta}(x_1|\vartheta)}{\sum_{\vartheta'} p_\Theta(\vartheta')p_{X_1|\Theta}(x_1|\vartheta')}$$

But what if we have n measurements to consider. Then we have $\Pr[\Theta = \vartheta|X_1 = x_1, X_2 = x_2, \dots]$ which can be computed by naive Bayes as:

$$= \frac{p_\Theta(\vartheta)\prod_{i=1}^n p_{X_i|\Theta}(x_i|\vartheta)}{\sum_{\vartheta'} p_\Theta(\vartheta')\prod_{i=1}^n p_{X_i|\Theta}(x_i|\vartheta')}$$

Also $p_{X|\Theta}(x_i, \dots, x_1|\Theta = \vartheta)$ is *joint PMF* and we assume conditional independence.

2.5 Bayes theorem using PDF

As for PMF we have Bayesian statistics for PDF.

$$f_{\Theta|X}(\vartheta|x) = \frac{f_\Theta(\vartheta) f_{X|\Theta}(x|\vartheta)}{\int_{-\infty}^{\infty} f_\Theta(\vartheta') f_{X|\Theta}(x|\vartheta') \, \mathrm{d}\vartheta'}$$

2.6 Beta distribution

To see some nice properties of Bayesian theorem we will look into one new distribution. We will have $\alpha, \beta \geq 1$ and $\vartheta \in [0,1]$. Then

$$f_\Theta(\vartheta) = \frac{\vartheta^{\alpha-1}(1-\vartheta)^{\beta-1}}{B(\alpha,\beta)}$$

Where $B(\alpha, \beta)$ is called **beta function** and for all α, β it is a constant. For example the beta function for $B(1,1)$ is equal to 1 from [0,1] and 0 otherwise. And $B(1,2) = \frac{1}{2}$. It serves as a normalizing constant for the beta distribution.

Firstly the maximum is at $\frac{\alpha-1}{\alpha+\beta-2}$ which is the **mode** (*cz: modus*).

Secondly:

$$B(\alpha, \beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-2)!} = \frac{1}{\binom{\alpha+\beta-2}{\alpha-1}}$$

Lastly $\mathbb{E}[\Theta] = \frac{\alpha}{\alpha+\beta}$ which is the **mean**.

Now we will look into the Bayesian theorem using Beta distribution as a prior and Binomial distribution as a likelihood.

$$p_{X|\Theta}(k|\vartheta) = \binom{n}{k} \vartheta^k (1-\vartheta)^{n-k}$$

$$f_{\Theta|X}(\vartheta|x) = c_1 \vartheta^{\alpha-1}(1-\vartheta)^{\beta-1} \cdot c_2 \vartheta^x (1-\beta)^{1-x} \cdot c_3 =$$

Where c_1,c_2,c_3 do not depend on ϑ and are some constants.

$$= c_4 \vartheta^{\alpha+k-1}(1-\vartheta)^{\beta+n-k-1}$$

And that is some other Beta distribution with $\alpha' = \alpha + x$ and $\beta' = \beta + n - x$. And also we have these point estimates:

1. MAP $\hat{\vartheta} = \frac{x}{n}$ which is same as likelihood.
2. LMS $\hat{\vartheta} = \mathbb{E}(\Theta|X = x) = \frac{x+1}{n+2}$

2.7 Normal random variable

Also we can look at Bayesian theorem with normal variables. *Note: This doesn't seem so interesting and useful, since it is only computation and nothing else.*

2.8 Conditional expectation

Firstly we will remind how expectation is defined. $\mathbb{E}[Y] = \sum_{y \in \mathrm{Img}(Y)} y \Pr[Y = y]$ if Y is discrete or $= \int_{-\infty}^{\infty} y f_Y(y)\, \mathrm{d}y$ if y is continuous. Now we will show how conditional expectation is defined.

$$\begin{aligned} \mathbb{E}[Y|A] &= \sum_{y \in \mathrm{Img}(Y)} y \Pr[Y = y|A] \\ &= \int_{-\infty}^{\infty} y f_{Y|A}(y)\, \mathrm{d}y \end{aligned}$$

Now if we have X,Y discrete random variables and $x \in \mathbb{R}$, then:

$$\mathbb{E}[Y|X = x] =: g(x)$$

So g is a function $\mathbb{R} \to \mathbb{R}$. Then

$$\mathbb{E}[Y|X] =: g(X)$$

So we have two functions $\Omega \to^X \mathbb{R} \to^g \mathbb{R}$. Now we will show one property which is called **Law of Iterated Expectation**.

$$\mathbb{E}[\mathbb{E}[Y|X]] =^{\mathrm{DEF}} \mathbb{E}[g(X)] =^{\mathrm{LOTUS}} \sum_{x \in \mathrm{Img}(X)} g(x) \Pr[X = x] =$$

$$=^{\mathrm{DEF}} \sum_{x \in \mathrm{Img}(X)} \Pr[X = x] \mathbb{E}[Y|X = x] = \mathbb{E}[Y]$$

Where the last equivalence is by the Law of total Expectation. So by this we get $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$ if $\mathbb{E}[Y] < \infty$.

Now, we will use a similar approach to find an alternative definition of variance.

Let $Y = \hat{Y} - \tilde{Y}$ where $\hat{Y}$ and $\tilde{Y}$ are statistically independent and $var(\tilde{Y}) = \mathbb{E}[\tilde{Y}^2]$

$$var(Y) = var(\hat{Y}) + var(\tilde{Y}) - 2cov(\hat{Y}, \tilde{Y})$$

From the property of the statistical independence we get $cov(\hat{Y}, \tilde{Y}) = 0$.

$$\mathbb{E}[(Y - E[Y|X])^2|X] = var[Y|X] =: h(X)$$

2.9 Law of iterated variance

$$\mathrm{var}[Y] = \mathbb{E}[\mathrm{var}[Y|X]] + \mathrm{var}[\mathbb{E}[Y|X]]$$

Or it is called an **Eve's rule** (*as E for expected value and V for variance*). We may simulate it by saying that the first part of the sum is expected value of variance within one group and the second part is inter group variance. *This is also partly from the example that was sadly omitted.*

Next we can show that Least Mean Square is iff condition expectation. That is for given Y what is the value of y that minimizes $\mathbb{E}[Y - y]^2$?

$$\mathbb{E}[Y - y]^2 = \mathbb{E}[Y^2] - 2y\mathbb{E}[Y] + y^2 = f(y)$$

$$f'(y) = -2\mathbb{E}[Y] + 2y = 0\mathbb{R}ightarrow y = \mathbb{E}[Y]$$

Now we want for all x find $y = y(x)$ such that $\mathbb{E}[(Y - y(x))^2|X = x]$ is minimized. We can show by similar calculation that $y(x) = \mathbb{E}[Y|X = x]$. And our best (in the LMS sense) estimation is $\hat{Y} = \mathbb{E}[Y|X]$.

3. Stochastic processes

Stochastic process is a sequence of random variables $X_1, X_2, X_3, \ldots$. We will show that there exist many of them.

- Markov chain (+ extra condition)
- Wiener process
  - Browner motion
  - Stock prices
  - Limit version of RN
- **Arrival times** or alternatively **waiting for success**.

We will be looking at the last type of the processes.

3.1 Bernoulli process (denoted as Bp(p))

That is we have $X_1, X_2, \ldots$ iid and each one of them is $X_i \sim \text{Ber}(p)$ so with probability p it is 1 and 0 with probability $1-p$.

Observation.
- $X_n, X_{n+1}, \ldots$ *is also Bp*(p)
- $X_N, X_{N+1}, \ldots$ *is also Bp*(p)*, with* N *a random variable dependent only on the past*

Then we will define $T = \min\{t : X_t = 1\}$ or by words the time of the first success / arrival. And we can easily see that $T \sim \text{Geom}(p)$ so $\mathbb{E}[T] = \frac{1}{p}$ and $\text{var}[T] = \frac{1-p}{p^2}$.

Now we will try to generalize this by T_k as the time of the k-th arrival. So $T_1 = T$. Or written as $T_k = \min\{t : X_1 + X_2 + \cdots + X_t = k\}$.

Other interesting variable is the k-th waiting time (inter arrival) and it will be denoted as L_k. To describe this variable it is the time between $k-1$ arrival and k-th arrival. Then it follows

$$L_k = T_k - T_{k-1} \text{ when we put } T_0 = 0$$

$$L_k \sim L_1 = T\mathbb{R}ightarrow L_k \sim \text{Geom}(p)$$

And all L_i are independent. From the other way we can define T_k as the sum $\sum_{i=1}^{k} = L_i$. So we can then get expected value and variance.

$$\mathbb{E}[T_k] = \mathbb{E}[L_1] + \mathbb{E}[L_2] + \cdots + \mathbb{E}[L_k] = \frac{k}{p}$$

$$\text{var}[T_k] = \text{var}[L_1] + \text{var}[L_2] + \cdots + \text{var}[L_k] = k \cdot \frac{1-p}{p^2}$$

How could we compute $\Pr[T_k = t]$ =? Easily we can compute this by convolution formula $\binom{t-1}{k-1} p^k (1-p)^{t-k}$.

Lastly we define N_t as the sum $X_1 + X_2 + \cdots + X_t$ which is the number of successes till the time t. And $N_t \sim \text{Bin}(t,p)$. So $\mathbb{E}[N_t] = tp$ and $\text{var}[N_t] = tp(1-p)$.

3.1.1 Alternative definition

We can define Bernoulli process by different definition. First we will define $L_1, L_2, \ldots$ as iid $\sim$ Geom(p) and then $T_k = \sum_{i=1}^{k} L_i$. And X_i is 1 if $T_k = i$ for some k or 0 otherwise. Then $(X_i)_i$ is Bp(p).

3.1.2 Merging of Bernoulli process

We will have two processes which are independent.

$$\begin{aligned} X_1, X_2, X_3, \ldots &\quad \mathrm{Bp}(p) \\ Y_1, Y_2, Y_3, \ldots &\quad \mathrm{Bp}(q) \end{aligned}$$

Then the merge is $Z_i = X_i$ or Y_i. Properly it is

$$Z_1, Z_2, Z_3, \ldots \quad \mathrm{Bp}(p + q - pq) = \mathrm{Bp}(1 - (1-p)(1-q))$$

3.1.3 Splitting Bernoulli process

We can also split one Bernoulli process. Firstly we got

$$Z_1, Z_2, Z_3 \ldots \quad \mathrm{Bp}(r)$$

If $Z_i = 1$ then $X_i = 1$ with probability α and 0 with probability $(1 - \alpha)$ and if $Z_i = 0$ then $X_i = 0$. By this construction we get new Bernoulli process.

$$X_1, X_2, X_3, \ldots \quad \mathrm{Bp}(\alpha r)$$

3.2 Poisson process (denoted as Pp(λ))

As we defined Bernoulli process we also can define Poisson process which can be described as a continuous approximation of Bp(p). Now the arrival times are real numbers.

Definition 12. *1. For any interval of length τ probability of k arrivals is the same. Denoted as $P(k, \tau)$.*

2. Number of arrivals in [a,b] is independent of number in $[0, a)$.

3. $P(0, \tau) = 1 - \lambda\tau + o(1)$, $P(1, \tau) = \lambda\tau + o(1)$, $P(k, \tau) = o(1)$. for $k \geq 2$ where $o(1)$ is something that goes to zero

Then the sequence $T_1.T_2, T_3, \ldots$ is Pp(λ) where Ts are the arrival times.

As in Bernoulli process we have T_k as the time of k-th arrival. Then N_T is the number of arrivals in [0,t] and $N_T \sim$ Pois(λt) so $P(k, t) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}$.

We can show that by the following approximation:

$$\begin{gathered} Pr[N_t = k] = P(k,t) \implies P(1, \tfrac{t}{l}) = \tfrac{\lambda t}{l} + o(1) \\ Pr[N_t = k] = P(k,t) \approx P[\text{there are K small intervals that has 1 arrival}] = \\ = Pr[Bin(l, \tfrac{\lambda t}{l}) = k] \qquad \implies lim_{l\to\infty} Bin(l, \tfrac{\lambda t}{l}) \to Pois(\lambda t) \end{gathered}$$

Then again $L_k = T_k - T_{k-1}$ so Pr[$L_k \geq t$] = Pr[no arrival in $[T_{k-1}, T_{k-1} + t]$] and that is equal to P(0,t) $= e^{-\lambda t}$. Next Pr[$L_k \leq t$] $= 1 - e^{-\lambda t}\mathbb{R}ightarrow L_k \sim$ Exp(λ).

3.2.1 Alternative description

As in Bernoulli process we can define Poisson process the other way around. We start with sequence of iid $L_1, L_2, \cdots \sim \text{Exp}(\lambda)$. Then T_k is the sum $T_k = \sum_{i=1}^{k} L_i$. And we also get the same N_t.

Theorem 6. *This also defines Pp(λ). In other words it satisfies all of the three properties.*

Again as in Bp we can see that expected value of T_k and variance is the sum of expected values (resp. variances) of L_i which are $\frac{1}{\lambda}$ (resp. $\frac{1}{\lambda^2}$). By convolution we get that

$$f_{T_k}(t) = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{(k-1)!}$$

3.2.2 Splitting of Pp

We have a $\text{Pp}(\lambda)$ and each one is split (1 or 0) with probability p (resp. $1-p$). And then we get two processes $\text{Pp}(p\lambda)$ and $\text{Pp}((1-p)\lambda)$ and these are independent. Two new processes have still the same properties but with new λ'. To properly show that this holds we need to show all the properties from the definition.

$$Pr[T_1 > t] = Pr[T > t \ \& \ T' > t] = \ldots$$

Remark. Proving independence is quite cumbersome. The proof is based on an example from the lecture.

3.2.3 Merging of Pp

If we have two processes $\text{Pp}(\lambda)$ and $\text{Pp}(\lambda')$ we can merge these to get $\text{Pp}(\lambda+\lambda')$. Again to properly show that this holds we must show that the min of two Exp distributions is again Exp distribution with the sum. Which is quite easy since they are independent, then we get the product of exponent functions which is the same as the sum of their exponents.

What if we look at the $\Pr[T-t > s|T > t]$ which by definition is $\frac{\Pr[T>s+t \wedge T>t]}{\Pr[T>t]}$ and that is equal to $\frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} e^{-\lambda s}$ and we get the property that the Poisson process is **memory-less** so it doesn't matter when we will start measuring our data.

4. Balls & Bins

This model is if we have m balls and n bins and for each ball we put it independently at random to one bin, where each bin has the same probability.

One well known problem is *Birthday paradox* where we have k people as balls and 365 days as bins. Then we are asking what is the probability that one bin has at least 2 balls.

$$Pr[\text{at least 2 balls in one bin}] = 1 - Pr[\text{max 1 ball in one bin}] =$$

$$= 1 - \prod_{i=1}^{m-1} \frac{n-i}{n} \approx 1 - \prod_{i=1}^{m-1} e^{\frac{-i}{n}} = 1 - e^{\frac{-m(m-1)}{2n}}$$

We also consider other properties, such as the expected number of empty bins:

$$Pr[\text{bin } i \text{ is empty}] = (1 - \frac{1}{n})^m \approx e^{\frac{-m}{n}}$$

$$\mathbb{E}[\#\text{ of empty bins}] = n(1 - \frac{1}{n})^m \approx ne^{\frac{-m}{n}}$$

Theorem 7 (Max Load Theorem)**.** *If $m = n$ and are big enough and $M = 3\frac{ln(n)}{ln(ln(n))}$ then $Pr[max \# of balls in a bin > M] < \frac{1}{n}$.*

Proof.

$$Pr[\text{bin \#1 has} \geq M \text{ balls}] \leq Pr[Bin(n, \frac{1}{n}) = M] < \frac{1}{M!} < (\frac{e}{M})^M$$

$$Pr[\text{any bin has} \geq M \text{ balls}] \leq$$

$$\leq Pr[\text{bin \#1 has} \geq M \text{ balls}] + \cdots + Pr[\text{bin \#}n \text{ has} \geq M \text{ balls}] \leq n(\frac{e}{M})^M$$

Now we will show that this expression is smaller than $\frac{1}{n}$. In order to do that, we wil take the logarithm of both sides and we get:

$$2ln(n) + M(1 - ln(M)) < 0$$

We will then substitute M and show that the inequality holds. □

M balls and bins have multiple applications. We will use it for hashing and sorting.

4.1 Bucket Sort Application

We want to sort $n = 2^k$ numbers from range $[0, 2^l - 1]$ where $l > k$. The numbers are uniformly random in this range.

4.1.1 Algorithm

1. Put input x to a bucket $b(x)$ where $b(x)$ is a hash function of x and bucket is a list.
2. Sort each bucket (list) by a bubble sort in quadratic time.
3. Merge the buckets.

Time Analysis

Parts 1 and 3 are linear in n. For part 2, we will consider $X_i = \#$of inputs in the ith bucket $\sim Bin(n, \frac{1}{n})$. Then $\mathbb{E}$ time $= \mathbb{E}\sum(c_i X_i^2)$. Finally, we will use the definition of variance to show that the expected time is $< 2cn$. Hence the whole algorithm has linear expected time.

4.2 Hash Collisions Application

We want to store n strings and search fast. Using the max load theorem, we will show that max running time with a big enough n is $< 3\frac{ln(n)}{ln(ln(n))}$.

Theorem 8. *Distribution of $X_1^{(m)}, \ldots, X_n^{(m)}$, where $X_i^{(m)}$ represents the number of balls in bin i, is the same as $Y_1^{(m)}, \ldots Y_n^{(m)}$, iid, where $Y_i^{(m)} \sim Pois(\frac{m}{n})$ and $\sum Y_i^{(m)} = k$*

Proof. It is based on the fact that $X_1^{(m)} \sim Bin(m, \frac{1}{n}) \approx Pois(\frac{m}{n})$ and then we show that $Pr[X_1^{(m)} = k_1, \ldots X_n^{(m)} = k_n] = P_x = P_Y = Pr[Y_1^{(m)} = k_1 \ldots | \sum Y_i = k]$ □

Theorem 9 (Max Load Theorem 2)**.** *If $m = n$ and are big enough and $M = \frac{ln(n)}{ln(ln(n))}$ then $Pr[max \# of balls in a bin < M] \leq \frac{1}{n}$.*

5. Non-parametric statistics

In parametric statistics, we assume that the data comes from a known distribution and we try to estimate a parameter of that distribution. In non-parametric statistics, we don't assume anything about the distribution of the data.

5.1 Permutation test

is a technique to decide whether observed random variables come from the same distribution or not.

$$X_1, \ldots, X_m$$

$$Y_1, \ldots, Y_n$$

H_0 : All of these random variables come from the same distribution.

The quantity computed from values in a sample (statistic) T is the difference of means of Xs and Ys.

$$T := \bar{X}_m - \bar{Y}_n$$

Alternatively we can use the two-sided test:

$$T := |\bar{X}_m - \bar{Y}_n|$$

We pick a paramater γ and if $T \geq \gamma$ then we reject H_0. In order to decide γ, we want our test to be statistically significant. Hence the following must hold:
$Pr[\text{wrong rejection}] < \alpha = 0.05$

So we will compute γ based on the set of all measured values. Next, the observations of groups X and Y are pooled, and the difference in sample means is calculated and recorded for every possible way of dividing the pooled values into two groups of size $|X|$ and $|Y|$. The set of these calculated differences is the exact distribution of possible differences under the null hypothesis that group labels are exchangeable.The p-value of the test is calculated as the proportion of sampled permutations where the difference in means was greater than T.

If $(m + n)!$ is too big, we can use a random permutation test. We will generate k random permutations and compute the test statistic for each of them.

5.2 One-sampled Sign test

$X_1, \ldots X_n$ *i.i.d.* They have unknown distribution which is continuous, has median μ, possibly mean μ and is symmetric around μ.

$H_0 : \mu = 0$

$Y_i = sgn(X_i)$ is either 1 or 0

$Y = \sum Y_i \sim Bin(n, \frac{1}{2})$ assuming H_0. Next, we consider the dsitrubution of Y and compute the quantiles based on α. If $Y > y_{1-\frac{\alpha}{2}}$ or $Y < y_{\frac{\alpha}{2}}$ then we reject H_0.

5.3 Paired Sign test

$(X_1, Y_2), \ldots, (X_n, Y_n)$
$H_0 : \mathbb{E}[X] = \mathbb{E}[Y]$ alternatively $\mathbb{E}[X - Y] = 0$
We create a new variable $Z_i = X_i - Y_i$ and we apply the one-sample sign test on Z_i.

5.4 Wilcoxon signed-rank test

The one-sample Wilcoxon signed-rank test can be used to test whether data comes from a symmetric population with a specified median.

$X_1, \ldots, X_n$ median is 0
$H_0 : \mu = 0$
We sort $|X_1|, \ldots, |X_n|$ and assign ranks $r_1, \ldots, r_n$ to them. In case the numbers are the same, we compute the mean of the range. Then we compute $T = \sum_{i=1}^{n} r_i sgn(X_i)$ which can be computed as $T = T^+ - T^-$
We reject the null hypothesis if T is too large or too small.

5.5 Mann-Whitney U-test

2-sample non-parametric test, which checks whether two samples come from the same distribution.

We compute statistics $U = \sum_i^{|X|} \sum_j^{|Y|} S(X_i, Y_j)$
where $S(X_i, Y_j) = 0$ if $X_i > Y_j$, $S(X_i, Y_j) = 1$ when it is the other way around and $\frac{1}{2}$ if they are equal.

It is a form of a permutation test.

5.6 Consequences of statistical designs

5.6.1 Simpson's paradox

Simpson's paradox is a phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined.

Example. Females at Harvard have overall smaller success rate than males. However, when compared their success rates in separate majors, females usually dominate. This means that most of the females apply to more competitive majors.

5.6.2 Time dependency

$X_1, \ldots X_n$ all tests assume *i.i.d.* however, in reality, the data is dependent. $\mathbb{E}[X_i]$ depends on i.

We can test this phenomenon by replacing X_i by $X_i - \mu$ where μ is the median of the measured data. Then we can observe the sequence of pluses and minuses. If the sequence is random, then we can assume that the data is independent.

6. Moment Generating Function

Definition 13. *If X is a random variable, $s \in \mathbb{R}$ then $M_X(s) = \mathbb{E}[e^{sX}]$ where M_X is the moment generating function of X.*

Theorem 10. *For all s where $M_X(s)$ is defined and finite:*

$$M_X(s) = \sum_{k=0}^{\infty} \frac{1}{k!}\mathbb{E}[X^k]s^k$$

Proof. $\mathbb{E}[X^k]$ is called the k-th moment. $\mathbb{E}[X^2] = var(X) + \mathbb{E}[X]^2$

$$e^s = \sum_{k=0}^{\infty} \frac{s^k}{k!}$$

$$\mathbb{E}[e^{sX}] = \mathbb{E}[\sum_{k=0}^{\infty} \frac{(sX)^k}{k!}] = \sum_{k=0}^{\infty} \frac{1}{k!}\mathbb{E}[X^k]s^k$$

□

For continuous distribution Y, we compute MGF with the help of LOTUS rule as follows:

$$M_Y(s) = \int_{-\infty}^{\infty} e^{sy} f_Y(y)dy$$

Theorem 11.

$$M_{aX+b}(s) = e^{bs}M_X(as)$$

Theorem 12. *if X and Y are independent, then $M_{X+Y}(s) = M_X(s)M_Y(s)$*

Theorem 13. $\exists \epsilon > 0 \forall s \in [-\epsilon, \epsilon] : M_X(s) = M_Y(s) \in \mathbb{R} \implies F_X = F_Y$

Theorem 14. $\exists \epsilon > 0 \forall s \in [-\epsilon, \epsilon] : M_{Y_n}(s) \to M_Z(s) \in \mathbb{R}$ & *F_Z is continuous* $\implies F_{Y_n} \to F_Z$

In this case instead of two random variables, we have a sequence of random variables.

Theorem 15 (Central Limit Theorem)**.** $X_1, \ldots, X_n$ *i.i.d.*, $\mathbb{E}[X_i] = \mu$, $var(X_i) = \sigma^2$, $Y_n = \frac{1}{\sigma\sqrt{n}}((\sum_{i=1}^n X_i) - n\mu)$. *Then Y_n converges to $\mathcal{N}(0, 1)$.*

Theorem 16 (Chernooff's theorem)**.** $X_1, \ldots, X_n$ *i.i.d.,*$\sim Bern(\frac{1}{2})$, $X = X_1 + \cdots + X_n$, $var(X) = n$, $t > 0 : Pr[X \leq -t] = Pr[X \geq t] \leq e^{\frac{-t^2}{2\sigma^2}}$.

6.1 Source coding theorem

How to encode the information in the most efficient way?
Model: sequence of $X_1, \ldots, X_n i.i.d.$ over finite alphabet.
Goal: find the most efficient encoding of the sequence.
$X = (X_1, \ldots, X_n)$
$L(n\epsilon) = min\{L : \exists C_n \subset A^n \ \ s.t.\ |C_n| < 2^l$ & $Pr[X \in C_n] \geq 1 - \epsilon\}$

Theorem 17 (Shannon's source coding theorem)**.**

$$\forall \epsilon > 0 : lim_{n\to\infty} \frac{L(n\epsilon)}{n} = H(X)$$