Why probability?

Phenomenon
↓
mathematical model

DETERMINISTIC → PROBABILISTIC

- parameters determine
- fixed outcome

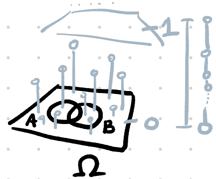
- parameters describe
- output is variable

Definitions

Sample space: set of all possible outcomes Ω

Event: subset of Ω, $A \in \Omega$

Probability P: is a function that assigns likelihood to events in Ω occurring



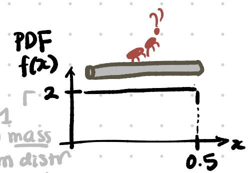
must satisfy the axioms

(1) $\sum_{A_i \in \Omega} P(A_i) = 1$ — something must happen

(2) $P(\emptyset) = 0$ — nothing never happens

(3) $P(A) + P(\bar{A}) = 1$ — either raining or not raining

(4) $P(A_i) \in [0,1]$ — mass is positive
finite $\Omega \rightarrow$ mass
infinite $\Omega \rightarrow$ density



≥ 1
- density, not mass
- for uniform distr
(# of points) P(point) = 1
∞ · ? = 1
↳ MUST BE ZERO
otherwise $\int \rightarrow \infty$

(5) If $A_1 \cdots A_n$ disjoint,
$A_i \cap A_j = \emptyset \quad \forall (i,j)$
$\sum P(\cup A_i) = \sum_i P(A_i)$

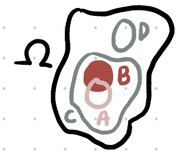
$\cup$: union (in either A or B)
$\cap$: intersection (in both A and B)

(6) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ — union

Independence: $P(A,B) = P(A)P(B)$ iff A & B are independent $A \perp\!\!\!\perp B$
$P(A,B)$: both events happening
arises from assumption or by construction
↓ coin flips $B \in B_1 \ldots 03$
↓ 6-dice: $A: \{2,4,6\}$, $B: \{1,2,3,4\}$

Conditional Probability: $P(A,B) = P(A|B)P(B)$

Suppose B has occurred.
This information changes probabilities of other events



$P(D|B) = 0$ $\quad (D, B \text{ disjoint}: D \cap B = \emptyset)$
$P(C|B) = 1$ $\quad (C \supset B)$
$P(A|B) = ???$ $\quad (\text{not disjoint}: A \cap B \neq \emptyset)$
given B has happened, what's prob. of A happening?
when do B & A overlap?

$$P(A|B) = \frac{\bullet}{\bullet} = \frac{P(A,B)}{P(B)}$$

What is $P(A|B)$ when $A \perp\!\!\!\perp B$ (they are independent?)

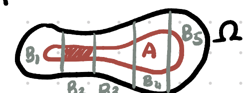
$P(A|B)$ shows up in

BAYES' LAW:
$P(A,B) = P(B,A)$
$P(A|B)P(B) = P(B|A)P(A)$
$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

- Bayes + billiards
- indep. thought of Laplace: P of cause given an event is proportional to P of event given the cause
- updates!

Marginalization:

1. partition Ω into B_i's: $B_1 \cdots B_n$

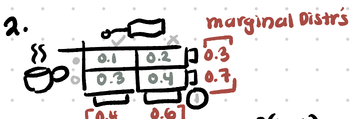


find $P(A)$ in terms of B_i's
B_i's disjoint $\cup P(AB) = \sum P(AB)$

$$P(A) = \sum_i P(A|B_i)P(B_i)$$

→ continuous analog

2.



$P(\circ|x) = \frac{P(\circ \cap x)}{P(x)} = \frac{0.2}{0.6}$

3. given joint distribution
$P(x) = \int \underbrace{P(x|y)P(y)}_{P(x,y)}\, dy$

Random variable: RV $\mathcal{X}: \Omega \to \mathbb{R}$ a very unfortunate name…

deterministic function that maps outcomes in sample space to real numbers.

So if Ω has a probability distribution over it, this induces a probability distribution of RV on $\mathbb{R}$

	Ω	$\mathbb{R}$	p( )
coins	H	→ 1	0.4
	T	→ 0	0.6

indicator R.V.

	Ω	$\mathbb{R}$	p( )
Distr. of sum of 2 dice	{1,1}	→ sum(1,1)	1/36
	⋮	⋮	⋮
	{6,6}	→ sum(6,6)	1/36

S: {i, j}
RV: i+j

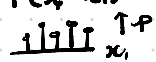
Random variables : MAPPING

Brownian motion

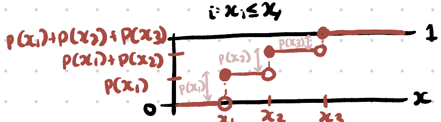
for $s \in \Omega$, $x \in \mathbb{R}$ RV: $s \to x$, can be

Discrete: x can take on finite # values

probability mass function: pmf: $p(x_i) = P(\mathcal{X} = x_i)$

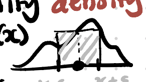


CDF: $F(x) = P(\mathcal{X} \le x) = \sum_{i: x_i \le x} p(x_i)$



continuous: x can take on infinite # values

probability density function
pdf $f(x)$



1. $f(x) \ge 0$
2. $\int_{-\infty}^{\infty} f(x) = 1$
3. $P(a \le X \le b) = \int_a^b f(x)dx$

$P(x = a) = 0 \quad \forall a \in \mathbb{R}$

$\mathbb{R}$
T_1 height(T_1)
T_2 height(T_2)
⋮ └ infinite Range

$F(x) = \int_{-\infty} f(x)dx$

$f(x) = \dfrac{dF(x)}{dx}$

$P(x-\varepsilon \le X \le x+\varepsilon) = \int_{x-\varepsilon}^{x+\varepsilon} f(x)dx \approx 2\varepsilon f(x)$

Cumulative distribution function of RV $\mathcal{X}$

$CDF(x) = F_{\mathcal{X}}(x) = \Pr(\mathcal{X} \le x)$

$\lim_{x \to -\infty} F_{\mathcal{X}}(x) = 0$

$\lim_{x \to +\infty} F_{\mathcal{X}}(x) = 1$

Expectation of RV $\mathcal{X}$ & function of $\mathcal{X}$ $g(\mathcal{X})$

	Discrete $\mathcal{X}$ (PMF)	continuous $\mathcal{X}$ (PDF)
$E[\mathcal{X}] =$	$\sum_i x_i p(x_i)$	$\int x f(x)dx$
$E[g(\mathcal{X})] =$	$\sum_i g(x_i) p(x_i)$	$\int g(x) f(x)dx$

$g(x)$: new random variable, want expectation

* statistical interpretation: $E[X] \approx \frac{1}{n}\sum_{i=1}^{n} x_i$
Law of Large Numbers

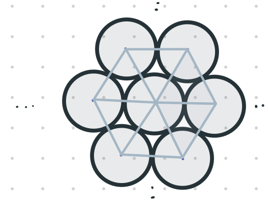
Repeated sampling $x_i = \mathcal{X}(s_i)$

* linearity: 2 rvs X & Y: $E(aX+bY) = aE(X)+bE(Y)$

Question: can 10 dots be covered with 10 identical coins with NO overlap???

area of plane covered by ○ : (tile w/ no gaps using hexagon)

$\dfrac{\text{area of ○ in ⬡}}{\text{area of ⬡}} : \dfrac{\pi r^2 + 6(\frac{1}{3}\pi r^2)}{\frac{3\sqrt{3}}{2}(2r)^2} = \dfrac{\pi}{2\sqrt{3}} = \boxed{0.9069}$



By linearity of expectation $10 \times 0.9069 \sim 9.069$

$E(\text{\# dots you cover}) = 9.069$
with this tiling

Since $E > 9$, there must be a tiling that can cover $\boxed{\text{ALL TEN}}$ points

Note that this approach does not work with 11 … $11(0.9069) = 9.98$ … so we are only promised coverage of 10/11 points by the above logic.

moment generating functions (MGF): n^{th} moment of RV $\mathcal{X}$

$$E[X^n] = \begin{cases} \sum_i x_i^n p(x_i) & \text{discrete} \\ \int x^n f(x)dx & \text{continuous} \end{cases}$$

moment	uncentered	centered	
1st	$E[X] = \mu$		avg
2nd	$E[X^2]$	$E[(X-\mu)^2]$	spread
3rd	$E[X^3]$	$E[(X-\mu)^3]$	asymm.
4th	$E[X^4]$	$E[(X-\mu)^4]$	tails

useful for studying $\sum$ RV's

*** Both CDF's & MGFs uniquely define a probability distribution ***

Storied Distributions

Discrete RV

1. Bernoulli RV: Two outcomes: success & failure

$$X = \begin{cases} 1 & \text{success} \quad \theta \\ 0 & \text{failure} \quad 1-\theta \end{cases}$$
(probability)

$X \sim \text{Bernoulli}(\theta)$ — distributed according to

$$p(x) : \begin{cases} \theta & \text{if } x=1 \\ 1-\theta & \text{if } x=0 \end{cases}$$

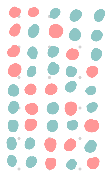
2. Binomial RV: independent Bernoullis w/ p success θ

$N = \#$ trials

$X = \#$ successes in trials $\sim \{0,1,2 \ldots N\}$

$X \sim \text{Binomial}(N, \theta)$

$$p(x) : \binom{N}{x} \theta^x (1-\theta)^{N-x}$$



$$\boxed{\binom{N}{x}} = \frac{N!}{x!(N-x)!}$$

ABCDE

$\left[\frac{}{5}\ \frac{}{4}\ \frac{}{3}\right]$

arrangements, order matters $= 5 \times 4 \times 3 = \frac{5!}{2!}$

But what if order doesn't matter? $\left(\frac{5!}{2!}\right)\frac{1}{3!}$ → shuffling chosen

3. Geometric RV: independent Bernoullis

$X = \#$ failures before success

$X \sim \text{Geom}(\theta)$

$$p(x) : (1-\theta)^x \theta$$

4. Poisson RV: essentially interested in # arrivals given rate of arrivals for memoryless process

mathematically it is a Binomial with λ finite successes

But N # trials $\to \infty$, θ p success $\to 0$ — $N\theta = \text{constant} = \lambda$

$x = \#$ arrivals

$\lambda = $ rate of arrivals

$X \sim \text{Poisson}(\lambda)$

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

*** Derivation ***

$$\binom{N}{x}\theta^x(1-\theta)^{N-x}$$

$$\lim_{\substack{N \to \infty \\ \theta \to 0}} \left[\frac{N!}{x!(N-x)!}\left(\frac{\lambda}{N}\right)^x\left(\frac{N-\lambda}{N}\right)^{N-x}\right] = \left[\underbrace{\frac{n(n-1)\cdots(n-x+1)}{n^x}}_{1} \cdot \frac{\overbrace{\left(1-\frac{\lambda}{n}\right)^n}^{e^{-\lambda}}}{\underbrace{\left(1-\frac{\lambda}{n}\right)^x}_{1}} \cdot \frac{\lambda^x}{x!}\right]$$

$e^{\cdot} = \lim_{n\to\infty}\left(1+\frac{\cdot}{n}\right)^n$

$$= \frac{\lambda^x}{x!}e^{-\lambda}$$

mailman week by week

A Poisson distribution & a Poisson process are not the same

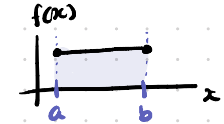
↳ see LHS of next page

Continuous RV

1. Uniform RV

$X \sim \text{Unif}(a,b)$

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$



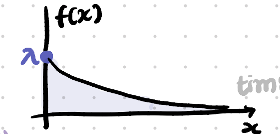
2. Exponential RV:

interarrival time of Poisson processes

λ: arrival rate (same λ in Poisson distr).

$X \sim \text{Expon}(\lambda)$

$$f(x) : \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$



Derivation: During interarrival, nothing arrives → Poisson w/ 0 arrivals in one unit of time

So $p(0) = \frac{\lambda^0}{0!}e^{-\lambda} = e^{-\lambda}$

$$P\left(\begin{matrix}\text{no arrivals} \\ \text{in time } t \text{ units}\end{matrix}\right) = P\left(\begin{matrix}\text{no arrivals} \\ \text{in time 0-1 unit}\end{matrix}\right) P\left(\begin{matrix}\text{no arrivals} \\ \text{in time 1-2 units}\end{matrix}\right) \cdots$$

$$= e^{-\lambda} e^{-\lambda} \cdots e^{-\lambda}$$

So in x time units,

$$P(X > x) = e^{-\lambda x}$$

Recognize CDF!!

$$1 - P(X \leq x) = e^{-\lambda x}$$

$$F(x) = 1 - e^{-\lambda x}$$

$$f(x) = \frac{d}{dx}F(x) = \boxed{\lambda e^{-\lambda x}} \quad \square$$

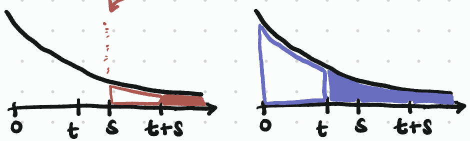
The exponential distribution is memoryless

Boundary for light/heavy tails.

Mathematical Encoding of Memorylessness

- Poisson processes are memoryless
- The exponential Distr. is the only memoryless continuous distr.
- A probability distribution is memoryless if

$$Pr(X > t+s \mid X > s) = Pr(X > t)$$



shifting by 's' makes no difference

$\boxed{P(X>t) = e^{-\lambda t}}$ recall from exp distr derivation

conditional probability definition

$$P(\underbrace{X>t+s}_{A} \mid \underbrace{X>s}_{B})\, P(\underbrace{X>s}_{\substack{B \\ e^{-\lambda s}}}) = P(\underbrace{X>t+s}_{\substack{A \\ e^{-\lambda(t+s)}}})$$

$$\boxed{P(X>t+s \mid X>s) = e^{-\lambda t}}$$

Poisson process

of ● → ∞

λ rate ●/●

# of a's : Poisson Distr.

↔ of ● btw ●'s : Expo. Distr.

↔ of ● btw α ●'s : Gamma Distr., $\alpha = \mathbb{Z}^+$

Heavy tail: tails heavier than the exponential

light tails ⟺ finite MGF ∀ orders

3. Gamma R.V.

waiting time for α arrivals of Poisson process

λ: rate of arrivals

α: # of arrivals

$$f(x; \lambda, \alpha) = \frac{1}{\Gamma(\alpha)} \frac{(\lambda x)^{\alpha}}{x} e^{-\lambda x}$$

Derivation:

$$P(X \le x) = 1 - \overbrace{P(X > x)}^{\text{same as before but cannot arrive } k=1,\ k=2,\ \vdots,\ k=k}$$

addition for mutually exclusive events

$$F(x) = 1 - \sum_{k=0}^{\alpha-1} \frac{(\lambda x)^k e^{-\lambda x}}{k!} \rightarrow \frac{dF(x)}{dx} = \frac{\lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}$$

↑ # of events

□

4. Gaussian Distribution RV (aka Normal)

$X_i \sim \mathcal{N}(\mu, \sigma^2)$

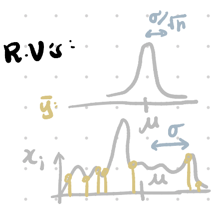
— very light tails

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

central limit theorem: Many quantities modeled by sums of R.V.s:

$$\approx \sum_{i=1}^{n} X_i \overset{\sim}{\sim} \mathcal{N}$$

└ CLT



If both 1st & 2nd moment defined, normalized sum of independent random variables w/ ANY underlying distribution approaches Normal. What does this mean?

Pick any $a, b \in \mathbb{R}$. $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $X_i \sim \mathcal{N}(\mu, \sigma)$

Concentration Bound:

$$\lim_{n\to\infty} Pr\left[a\frac{\sigma}{\sqrt{n}} \le \bar{y} - \mu \le b\frac{\sigma}{\sqrt{n}}\right] = \underbrace{\frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}t^2}\, dt}_{\text{PDF}} : \bar{y} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

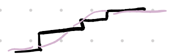
Parameter Estimation

Statistical functional $T: F \to \mathbb{R}$

can define θ, parameter of distribution as a functional $\theta = T(F)$

Ex: means variances medians

approximate CDF with ECDF



$\downarrow d/dx$ $\quad$ $\downarrow d/dx$

$f(x)$ $\quad$ $\hat{f}(x)$ $\frac{1}{n}\sum_{i=1}^{n} \delta(x-x_i)$

Estimate $\hat{\theta} = T(\hat{F})$ plug in empirical data

$\hat{\theta}$ is called a plug-in estimate

$\hat{\theta}$'s have biases: $\langle\hat{\theta}\rangle - \theta = \int \hat{\theta} f(x)\,dx - T(F)$

How off is this estimate on average?

Question: Given a set of data, we can estimate parameters of interest via plug-in estimates, but how do we account for sampling variation?

Solution: Bootstrapping! → confidence intervals!

sampling your dataset with replacement

95% of the time, a 95% interval of $\hat{\theta}$ will contain θ

Maximum Likelihood Estimate

$\vec{y} = \{y_1, y_2, y_3, \cdots y_n)$

Likelihood: $L(\theta; \vec{y}) = f(\vec{y}; \theta)$

for iid, $L(\theta; \vec{y}) = \prod_{i=1} f(y_i)$ sums are nicer than products

log-likelihood $\ell(\theta; \vec{y}) = \log L(\theta; \vec{y}) = \sum_i \log f(y_i)$

$\theta_{MLE} = \text{argmax}_\theta\, \ell(\theta; \vec{y})$ set $\frac{\partial \ell}{\partial \theta} = 0$

logarithm is monotonic, order is preserved

(A) constructing conf ints non-parametrically

$\{x_1\ x_2\ x_3\ x_4\}$ $\quad \theta$

$\{1, 2, 3, 4\} \to \hat{\theta}$ plug in estimate

Bootstrap samples x^*

$\{2, 2, 4, 3\}$
$\{1, 4, 3, 1\}$
$\vdots$
$\{2, 3, 1, 1\}$

N_samples

$\to$ $\hat{\theta}^*_1$, $\vdots$, $\hat{\theta}^*_{N\_samples}$

$\hat{\theta}^*$ Bootstrap Replicates

for αth conf-ints, retrieve percentiles of $\hat{\theta}^*$

(B) Constructing conf. ints parametrically

$\{x_1\ x_2\ x_3\ x_4\}$

$\{1, 2, 3, 4\} \xrightarrow{MLE} \theta^*_{MLE}$ maximum likelihood estimate

Bootstrap samples x^*

$\{2, 2, 4, 3\}$
$\{1, 4, 3, 1\}$
$\vdots$
$\{2, 3, 1, 1\}$

N_samples

$\xrightarrow{MLE}$ $\theta^*_{1,MLE}$, $\vdots$, $\theta^*_{N,MLE}$

for αth conf-ints, retrieve percentiles of $\hat{\theta}^*$

- Sometimes MLE = plug-in, but this is not always the case
- parametric inference assumes the model you have is true, and then we optimize under that assumption

1. Prof. Leonard Schulman's CS 150 2018 Lecture Notes
2. Prof. Kostia Zuev's ACM 116 Course Webpage