Rapid Review of Probability

Example: Coin toss : two outcomes, H+T

We say: "prob of H = prob of T = $\frac{1}{2}$".

Why? There is nothing "random" about a coin toss.

↳ If I can toss it in exactly the same way under the exact same conditions, the outcome will be the same.

↳ Describing a coin toss is very difficult.

→ Probability is a good "model" for situations that are difficult and for which we are willing to say that we don't "know" everything.

→ We need to formalize this intuition.

Definition "Sample Space" Ω : the space of all outcomes of a (well defined) experiment.

Examples:

↳ Experiment: Coin toss: outcomes are H+T. (I)

↳ Experiment: "Wait for tomorrow and observe the weather"
$\Omega = \{$all possible weather situations$\}$ (II)

↳ Experiment: "Wait for tomorrow and measure temperature at 12^{00} on top of ENR2" (III)
$\Omega = \{$All possible temperatures$\}$

Definition: An "event" is a subset of Ω.

Examples: * for Exp I: ⇝ an event is T

⇝ another event is H

* for Exp II: ⇝ "a sunny day with $T < 102°F$"

* for Exp III: ⇝ T is less than 108°F

⇝ $92°F \le T \le 108°F$

Definition σ-algebra.

The set of all events we want to consider, $\mathcal{B}$, must be a σ-algebra.

That is, $\mathcal{B}$ must satisfy:

(i) $\phi \in \mathcal{B}$

↑ "the empty set"

(ii) if $A \in \mathcal{B}$, then $CA \in \mathcal{B}$

↑ "Complement of A", i.e., everything that is not A.

(iii) If $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$ is a collection of sets, then the unions of all elements of $\mathcal{A}$ must also be in $\mathcal{B}$.

Examples

(a) Coin toss:

$\Omega = \{H, T\}$

$$\underbrace{\mathcal{D} = \{\phi, \{H,T\}\}}$$

describes that "nothing happened" (ϕ) or "something happened" (either H or T).

$$\underbrace{\mathcal{D} = \{\phi, H, T\}}$$

describes what the outcome of the experiment "toss a coin" is.

(b) Dice

$\Omega = \{1, 2, 3, 4, 5, 6\}$

$$\underbrace{\mathcal{D} = \{\{1,3,5\}, \{2,4,6\}, \phi\}}$$

outcome is odd or even

$$\underbrace{\mathcal{D} = \{1, 2, 3, 4, 5, 6, \phi\}}$$

a certain number is up after tossing the dice.

Definition: A "probability measure" is a function

$$P : \mathcal{B} \to \mathbb{R}$$

($\mathbb{R}$ the real numbers)

that satisfies:

(i) $P(\Omega) = 1$

(ii) $0 \le P \le 1$

(iii) Let $\{A_1, A_2, \dots, A_n\}$ be a finite collection of events $A_i \in \mathcal{B}$ and $A_i \cap A_j = \phi$ (*) for $i \neq j$. Then $\underbrace{P(\bigcup_i A_i) = \sum P(A_i)}$

Probability of union of sets = Sum of Probabilities

Definition: The triple $(\Omega, \mathcal{B}, P)$ is a probability space

Ω ~ all possible outcomes of an experiment

$\mathcal{B}$ ~ outcomes we are interested in

P ~ probabilities that outcomes in $\mathcal{B}$ occur

$P(A) = 0$ means A almost never occurs

$P(A) = 1$ means A almost always occurs

Example: Coin toss

$$\left.\begin{aligned} \Omega &= \{H, T\} \\ \mathcal{B} &= \{\phi, H, T\} \\ P(H) &= 1/2 \\ P(T) &= 1/2 \end{aligned}\right\}$$ This is a model that describes the experiment "toss a coin".

(*) $A_i \cap A_j = \phi$ means intersection of sets is empty.

Is this a good model?

- The model says that in about half of your tries, you will see a H, you will see T in the other half
- Does this agree with what we observe?

  Find out in HW!

  Toss a coin 10 times. How often do you see H?
  Toss a coin 100 times. How often do you see H?
  Toss a coin 500 times. How often do you see H?

  use "parallel experiments" for the last questions, i.e., make 5 people toss a coin 100 times.

From now on, assume that $\Omega = \mathbb{R}$ (all real numbers)

Definition: A "random variable" is a function $X : \Omega \to \mathbb{R}$ such that

$$A_\eta = \{\omega \in \Omega : X(\omega) \le \eta\}$$

is an element of $\mathcal{B}$ for all η.

- This means that we can assign probabilities to all events $X(\omega) \le \eta$ (the r.v. is less than some number)
- Loosely speaking, a r.v. is a variable whose value is not known, but it is possible to assign probabilities to the occurrence of various values.
- A r.v. takes a probability space and maps it to observables we care about.

<u>Definition</u> The "probability distribution function" of a r.v. X is

$$F_X(\eta) = P(\{\omega \in \Omega : X(\omega) \le \eta\}).$$

The probability distribution function (PDF) has the following properties:

(i) $F_X(-\infty) = 0$

(ii) $F_X(\infty) = 1$

(iii) $F_X(\eta_1) \ge F_X(\eta_2)$ if $\eta_1 \ge \eta_2$

(non-decreasing fctn)

<u>Example:</u> Dice:

$\Omega = \{1, 2, 3, 4, 5, 6\}$

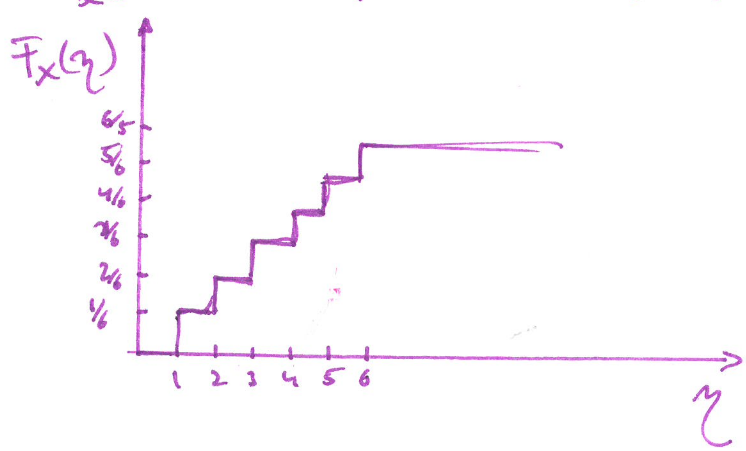
$\mathcal{B} = \{\emptyset, 1, 2, 3, 4, 5, 6\}$ (we care about what # is up)

$P(i) = 1/6$ for $i = 1, 2, \ldots, 6$

$$F_X(0) = P(\{\omega \in \Omega : X(\omega) \le 0\}) = 0$$

$$F_X(1-\varepsilon) = P(\{\omega \in \Omega : X(\omega) \le 1-\varepsilon\}) = 0 \quad \text{for all } \varepsilon > 0$$

$$F_X(1+\varepsilon) = P(\{\omega \in \Omega : X(\omega) \le 1+\varepsilon\}) = 1/6 \quad \text{for } \varepsilon \ge 0,\ \varepsilon < 1$$



Definition: If $F_X'(\eta)$ exists, then $p_X(\eta) = F_X'(\eta)$ is the "probability density function" (pdf) of $\underset{\text{r.v.}}{}$ X.

- For the rest of this class, we consider only r.v.'s for which $F_X'(\eta)$ exists.
- We will call the pdf of r.v. X also the (probability) distribution of X.

Properties of the pdf:

(1) Since $F_X(\eta)$ is non-decreasing $p_X(\eta) \geq 0$.

(2) $F(\eta + d\eta) - F(\eta) = p(\eta)\, d\eta$

$P(\{\omega \in \Omega : \eta < X(\omega) < \eta + d\eta\})$

$=$ pdf $\times$ Volume

(3) $F_X(\eta) = \int_{-\infty}^{\eta} p(y)\, dy$

(4) $\int_{-\infty}^{\infty} p(y)\, dy = F_X(\infty) - F_X(-\infty) = 1$

[redacted]

Multivariate r.v.

Probability space $(\Omega, \mathcal{B}, P)$

$$X : \Omega \to \mathbb{R}^n$$

"Components of X are r.v."

$$X_1 : \Omega \to \mathbb{R}$$
$$X_2 : \Omega \to \mathbb{R}$$
$$\vdots$$

$$A_\eta = \{\omega \in \Omega : X_1(\omega) \le \eta_1,\ X_2(\omega) \le \eta_2 \dots \}$$

$\uparrow$ is a element of $\mathcal{B}$ for all η

$\eta = (\eta_1, \eta_2, \dots, \eta_n)$ a vector

Probability distribution fcn: $F_X(\eta) = P(A_\eta)$

If $F_X(\eta)$ is differentiable, $\frac{\partial^n}{\partial \eta_1 \cdots \partial \eta_n} F_X(\eta) = p(x)$

is the probability density of X.

<u>Note:</u> Probability space itself usually not really known & not important.
What is known are <u>assumed</u> probability measures and probability densities.

Examples of r.v.s

(1) Uniformly distributed r.v.

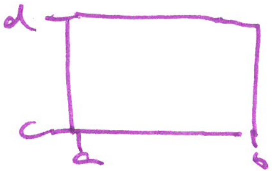
$X \sim \mathcal{U}[a, b]$ means every value of x in $[a, b]$ has the same probability.

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

(2) Multivariate uniform r.v.

$X \sim \mathcal{U}$ means every value in a cube has the same probability.

$p(x) \propto \text{Const.}$

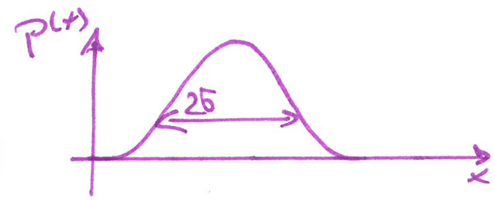


in 2D: $p(x) \propto \frac{1}{(b-a)(d-c)}$.

(3) Gaussian r.v. ("normal distribution")

$X \sim \mathcal{N}(\mu, \sigma^2)$, $\sigma > 0$

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

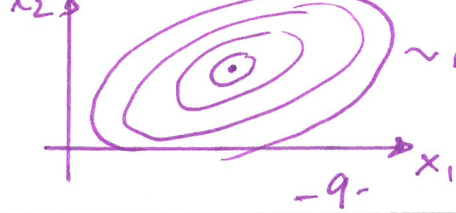


~ x is likely to be within $[\mu - 2\sigma, \mu + 2\sigma]$ (see later about why)

(4) Multivariate Gaussian

$X \sim \mathcal{N}(\mu, \Sigma)$ $\Sigma > 0$, $\Sigma^T = \Sigma$

$$p(x) = (2\pi)^{-n/2} |\Sigma|^{1/2} \exp\left(-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$



~ level sets of p.

⑤ χ-distribution

$$y = \left(\sum_{i=1}^{k} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \right)^{1/2}$$

$k > 0$ - degrees of freedom
Note that $y \geq 0$

$x_i \sim N(\mu_i, \sigma_i^2)$

$$P(y) \propto 2^{1-k/2} \, y^{k-1} \, e^{-y^2/2}$$

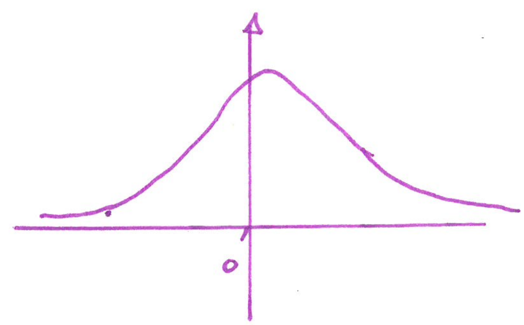
⑥ χ^2-distribution

$$y = \sum_{i=1}^{k} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 , \quad x_i \sim N(\mu_i, \sigma_i^2)$$

$$P(y) \propto y^{k/2 - 1} \, e^{-y/2} , \quad y \geq 0$$

⑦ Students t-distribution

$$p(x) \propto \left(1 + \frac{x^2}{\nu} \right)^{-\frac{\nu+1}{2}}$$



"looks like Gaussian but has "heavier tails"".

⑧ Multivariate t-distribution

$$P(y) \propto \left[1 + \tfrac{1}{\nu} (x-\mu) \Sigma^{-1} (x-\mu) \right]^{-\frac{\nu+p}{2}}$$

p = dim (x)

Constructed as: $x \sim N(0, \Sigma)$, $u \sim \chi^2_\nu$

$$y = \left(\frac{\nu}{u} \right) \cdot (x - \mu)$$

is (Student) t distributed

References:

Chan & Hald, 2013, Chapter

Read & Cotter, 2015, Chapter

Homework:

(1) Is $p(H) = p(T) = \frac{1}{2}$ a good model for a coin toss? How often do you observe H when you toss a coin 10, 50, 100, 500 times?

(2) For Gaussian distribution $p(x) = N(\mu, \sigma^2)$, why is $\sigma > 0$?
For multivariate Gaussian $p(x) = N(\mu, \Sigma)$ why is $\Sigma > 0$? why is $\Sigma^T = \Sigma$?