LMS Statistics and hypothesis testing

November 2023

LMS statistics and hypothesis testing

Chapter 1. Random events and probability theory

i) What is probability
ii) Discrete distributions
iii) Continuous distributions

Chapter 2. Confidence intervals and central limit theorem

i) Confidence intervals
ii) The Central Limit theorem

Chapter 3. Hypothesis testing

i) Formulate hypothesis H_0, H_1
ii) Quantify significance. p-value
iii) Comparing means. t-test.

1.1 Random events and probability

* Random events ("stochastic")

  Something whose output we don't know

* Probability : number $\in [0,1]$ quantifying certainty / "surprise".

  Example : tossing coins (H, T)

  $P(H) = 0 \rightarrow$ Certain I will never get H

  $P(H) = 1 \rightarrow$ " always get H

  $0 < P(H) < 1 \rightarrow \exists$ level of uncertainty / "surprise"

* Unitarity : the sum of probabilities for all possible outcomes must add up to 1.

  Example : tossing coins

  $P(H) + P(T) = \frac{1}{2} + \frac{1}{2} = 1$ ✓

1.2 Discrete probability distributions

* Discrete: number of possible outcomes is a finite number. { coins, dice, counting

i) Binomial distribution

"How many times X I get a specific result in n trials, if the probability of each success is p"

$$B(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}$$

x : number of successes
n : number of trials
p : probability each success.

Example: Probability of 5 times H tossing 10 times a coin

$$B\left(5;10,\frac{1}{2}\right) = \binom{10}{5}\left(\frac{1}{2}\right)^5\left(1-\frac{1}{2}\right)^{(10-5)} = 0.246$$

Exercise

Example: Probability of 3 times a 6 in 10 dice

$$B\left(3;10,\frac{1}{6}\right) = \binom{10}{3}\left(\frac{1}{6}\right)^3\left(1-\frac{1}{6}\right)^{(10-3)} = 0.155$$

Exercise

Example: Probability of passing exam (A, B, C) of 10 quest.

$$B\left(5;10,\frac{1}{3}\right) = \binom{10}{5}\frac{1}{3}^5\left(1-\frac{1}{3}\right)^5 = 0.137$$

Exercise

iii) Poisson distribution

"How many times I observe an event in an interval, knowing λ"

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

x : number of times I observe

λ : observed average.

Example: Probability of observing 3 cancer patients, with $\lambda = 5$.

$$P(3;5) = \frac{e^{-5}\,5^{3}}{3!} = 0.140$$

Exercise

Example: Probability of observing 5 or less patients in same λ.

$$P(x \leq 5; 5) = P(1) + P(2) + \ldots + P(5)$$

$$= \underbrace{\sum_{x=1} P(5; \lambda = 5)}_{\text{Cumulative distribution function } \underline{\text{cdf}}} = 0.616$$

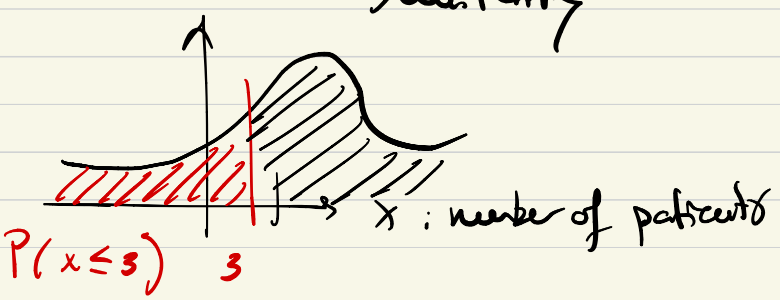
Exercise

Example: Probability of more than 5

$$P(x > 5, \lambda = 5) = 1 - \text{cdf}(5) = 0.394$$

Exercise.

↑ unitarity



1.3 Continuous distributions

* Continuous : amount of possible outcomes is infinite / uncountable

Coins : 2 outcomes $(H, T) \rightarrow p = \frac{1}{2}$

Dice : 6 outcomes $(1,2,3,4,5,6) \rightarrow p = \frac{1}{6}$

Discrete cases "Mass dist."

Frequentist definition ✓

Continuous variable (T, h, conc.) : ∞ outcomes $\longrightarrow p = \frac{1}{\infty} = 0$ WTF

↳ Frequentist approach does not work

Need a new mathematical object "Density distribution"

i) Discrete case

Probability $P \in [0, 1]$

Unitarity $\sum_{\forall x_i} P(x_i) = 1$

ii) Continuous case

Density $f(x)$

Unitarity $$\int_{-\infty}^{\infty} dx\, f(x) = 1$$

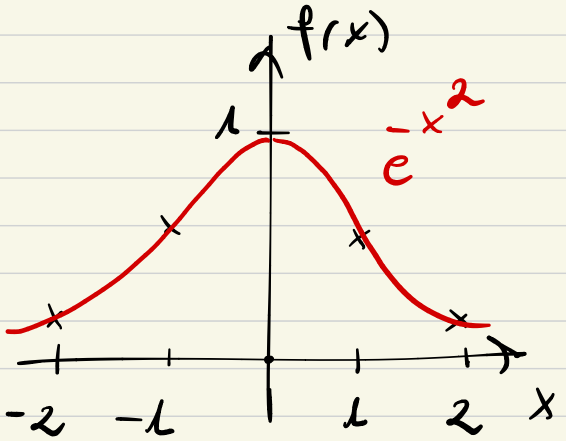
* Gaussian

i) Gaussian distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad \begin{cases} \mu : \text{mean value} \\ \sigma : \text{standard deviation} \end{cases}$$

* Consider simplest case ($\mu = 0$; $\sigma = 1$; normalization factor 1)

$f(x) = e^{-x^2}$; plot some values



$x = 0$; $f(0) = e^{-0} = 1$

$x = 1$; $f(1) = e^{-1} = \frac{1}{e} = 0.37$

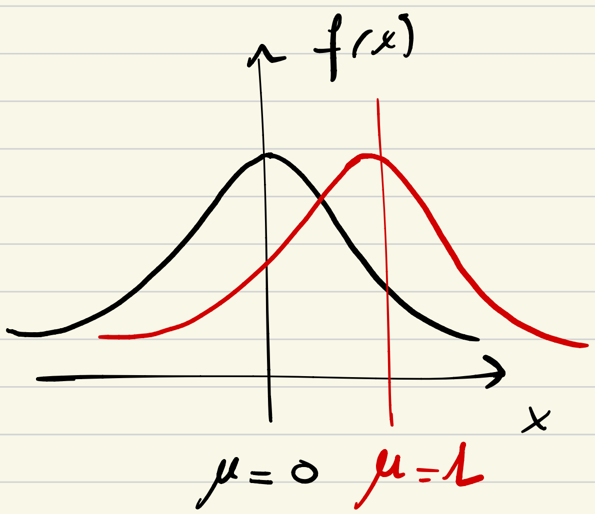
$x = -1$; $f(-1) = e^{-(-1)^2} = \frac{1}{e} = 0.37$

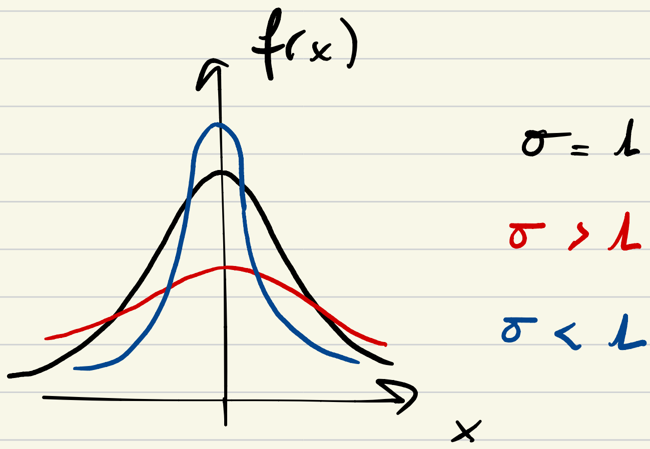
$x = \pm 2$; $f(\pm 2) = e^{-4} = \frac{1}{e^4} = 0.018$

* Switch back on μ and σ

$$f(x) = e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$\begin{cases} \mu \text{ will displace the central value} \\ \sigma \text{ will drive sharper/wider shape} \end{cases}$





* Add normalization factor

i) Remember discrete probability: Unitarity $\sum_{\forall x_i} P(x_i) = 1$

ii) Continuous case. Density distributions $\int_{-\infty}^{+\infty} dx\, f(x) = 1$

$$\int_{-\infty}^{\infty} dx\; e^{-\frac{1}{2\sigma^2}(x-\mu)^2} = \sigma\sqrt{2\pi}$$

Exercise

* for $f(x)$ to behave as a probability, we add a normalization factor

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

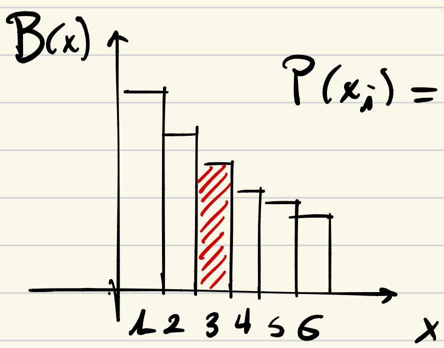
Gaussian distribution; $\int_{-\infty}^{\infty} dx\, f(x) = 1$

* Particular case $\mu = 0$; $\sigma = 1$

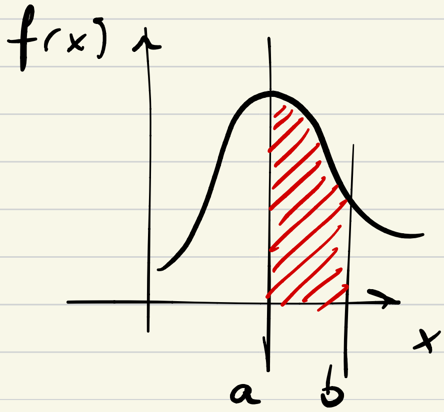
$$\left\{ f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2} \right\}$$

Normal distribution $N(0,1)$

* In continuous distributions, we can only compute probability in a given range. Never a single value.



$P(x_i) = B(x_i)$ ✓



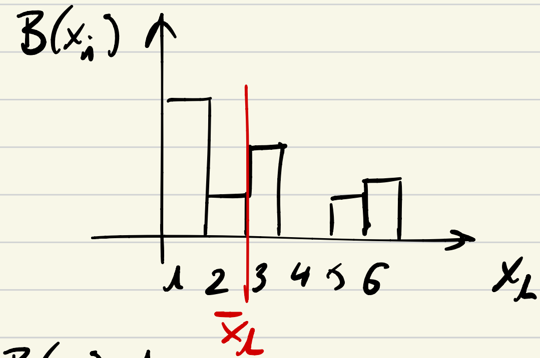
$P(a < x < b) =$

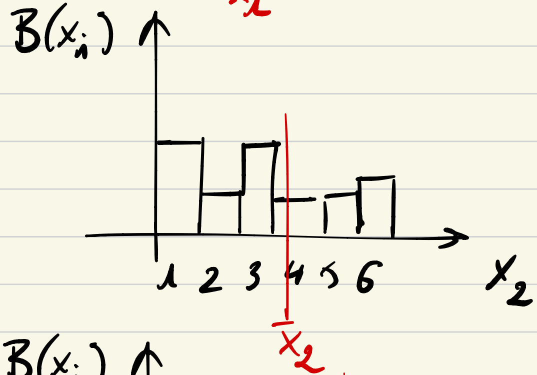
$cdf(b) - cdf(a)$

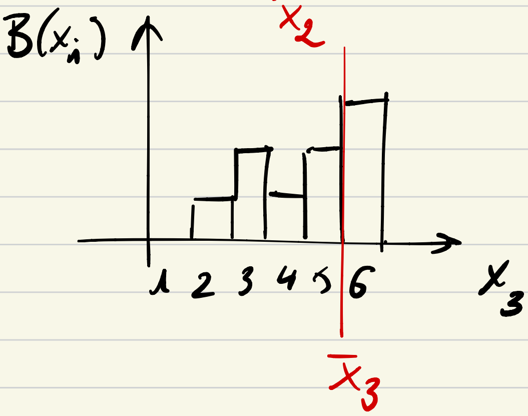
2.1 The Central Limit Theorem

"For a set of n independent and identically distributed (iid) random variables x_i the sample means tends to a gaussian distribution for large values of n, regardless of how x_i are themselves distributed"

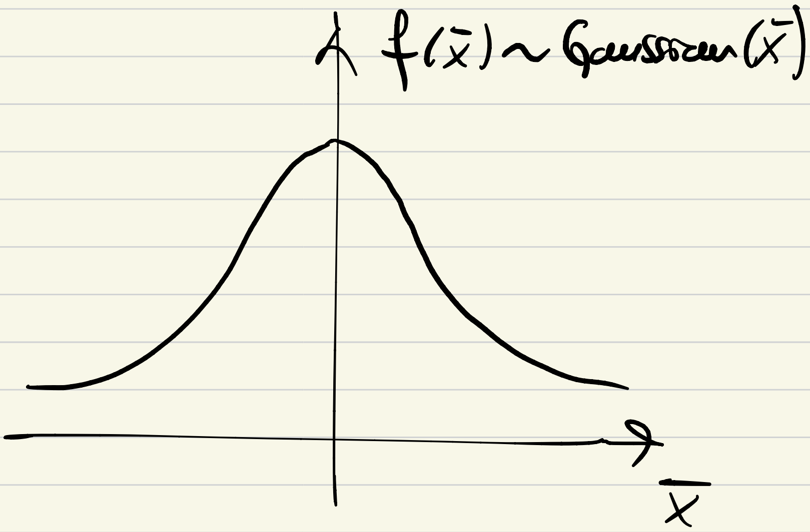
i) Consider 10 rolls of a dice. How many times I observe a 6.
Binomial distributed $B(x; 10, 1/6)$. Repeat n times.







* As the number of repetitions n increase (↑ sample size), the distribution of means $\bar{x}_i$ follows a gaussian distribution



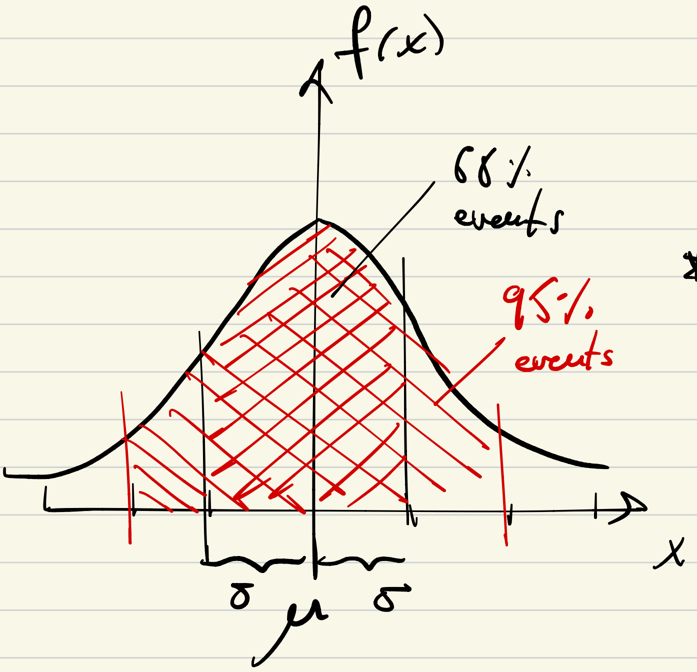
ii) Build new random variable
$\bar{x} = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n\}$

* Mean value μ of the $\bar{x}_i$ variable
* Standard error (SE)

$$SE = \frac{\sigma_i}{\sqrt{n}}$$ as I increase n, ↓SE

2.2 Confidence Intervals.

i) Consider a Gaussian distribution $f(x; \mu, \sigma)$



* 68 % of events are contained in the $(\mu-\sigma, \mu+\sigma)$ interval. 1σ conf. int.
* 95 % of events are contained in the $(\mu-2\sigma, \mu+2\sigma)$ interval. 2σ conf. int.
* 99 % of events are contained in the $(\mu-3\sigma, \mu+3\sigma)$ interval. 3σ conf. int.

ii) Any random variable can be normalized / "standardized" just by subtracting the mean and dividing by the standard deviation.

x: random distributed; same μ, σ.

standardized x; $\left\} z = \dfrac{x-\mu}{\sigma} \right\{$ follows a $N(0,1)$ distribution, also referred to as z score or z distribution.

iii) We want to estimate the mean μ

The best estimator is the sample mean $\bar{x}$, or "average"

Normally, we want to build

$$z = \frac{\bar{x}-\mu}{s/\sqrt{n}} \quad ; \text{such that} \quad P(\bar{x}-z < \mu < \bar{x}+z) = P_{\text{critical value}}$$

③ Hypothesis testing

i) Check if an observation is compatible with a given hypothesis H_0

* statistic test : a function / number computed out of our data
  (t-test, χ^2 test, Fisher test, Wald test, ...)

* p-value : probability of obtaining a result at least as extreme as the one actually observed, under the assumption that our null hypothesis H_0 was correct.
  (how likely / unlikely was to observe this result)

i) ↑ p-value : it was likely to obtain this result. Accept H_0.
ii) ↓ p-value : " unlikely " . Reject H_0.

ii) General approach

1. Formulate null hypothesis H_0 and significance level α.
2. Make measurement / observation.
3. Compute statistic test
4. Compute p-value
5. If p-value $< \alpha$, reject H_0. Otherwise, accept H_0.

Example: Check if a die is biased.

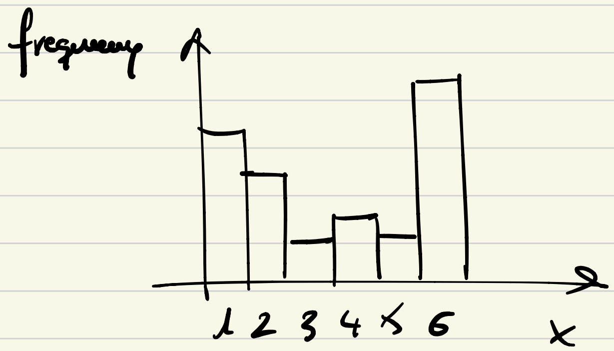
* You suspect your opponent is using a loaded die

i) Formulate null hypothesis $H_0 : p = \frac{1}{6}$

// alternative // $H_1 : p > \frac{1}{6}$

Choose significance level $\alpha = 0.01$

ii) Collect data. Roll die 100 times



43 times out of 100 we get a 6.

iii) Compute probability of 43 or less 6s, given H_0.
Given by Binomial distribution with $p = 1/6$

$$B_{cdf}(43; 100, \tfrac{1}{6}) \sim 1$$

* Exercise

iv) Compute p-value. Probability of obtaining at least a value as extreme as 43.

$$\text{p-value} = 1 - B_{cdf}(43; 100, \tfrac{1}{6}) = 10^{-10}$$

$\hookrightarrow$ p-value $< \alpha$. $\exists$ evidence to reject H_0.

Example: check if coin is fair

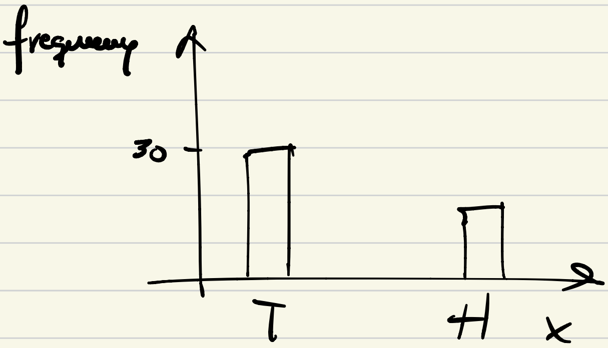
* You suspect your opponent is using a biased coin

i) Formulate null hypothesis $H_0: p = \frac{1}{2}$

" alternative " $H_1: p \neq \frac{1}{2}$

Choose significance level $\alpha = 0.05$

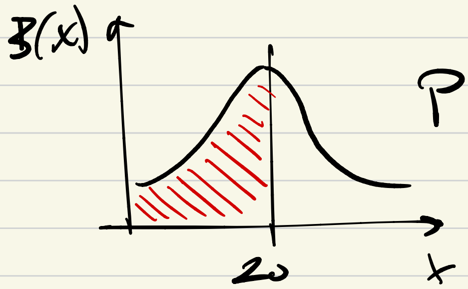
ii) Collect data. Roll die 100 times



20 heads in 50 flips of a coin

iii) Compute probability of 43 or less 6s, given H_0.

Given by Binomial distribution with $p = 1/6$



$P(x=20) = B_{cdf}(20; 50, \frac{1}{2}) = 0.1$ *

iv) Compute p-value.

Now our H_1 is just $p \neq \frac{1}{2}$, either larger or smaller.

p-value $= 2 \cdot cdf(20) = 0.2 < \alpha$ ∄ evidence to reject H_0.