RCDS Statistics I - Imperial College London

Introduction to random sampling and hypothesis testing

Jesús Mirasso Elizari

November 2023

Random sampling 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) Z and ϵ distribution

iii) 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 x_i must add up to 1.

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

Example: tossing coins

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

Example: rolling dice

$$P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = \frac{1}{6} + \frac{1}{6} + \dots + \frac{1}{6} = 1 \quad \checkmark$$

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 # 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.

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

(*) Exercise

ii) Poisson distribution

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

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

x: number of times I observe

λ: observed average.

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

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

Excercise

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

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

$$= \sum_{x=1}^{5} P(x;\lambda=5) = 0.616$$

Excercise

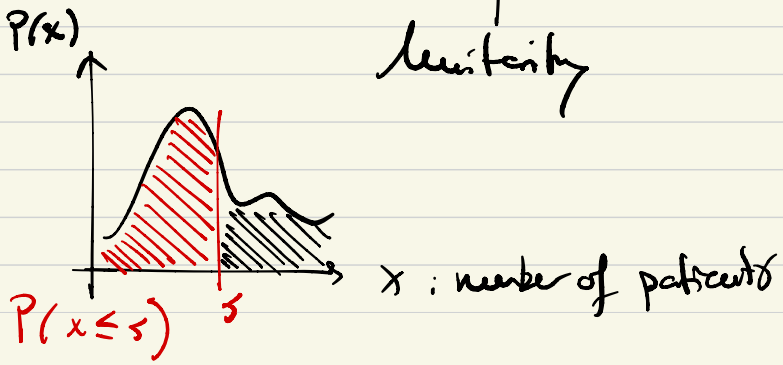
Cumulative distribution function cdf

Example: Probability of more than 5

$$P(x>5, \lambda=5) = 1 - cdf(5) = 0.392$$

Excercise

Unitarity



1.3 Continuous distributions

* Continuous : amount of possible outcomes is infinite / uncountable

Coin: 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 $(T, h, \text{conc.})$: ∞ outcomes $\rightarrow p = \frac{1}{\infty} = 0$ 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

"Random variable x centered at some μ and spreaded σ"

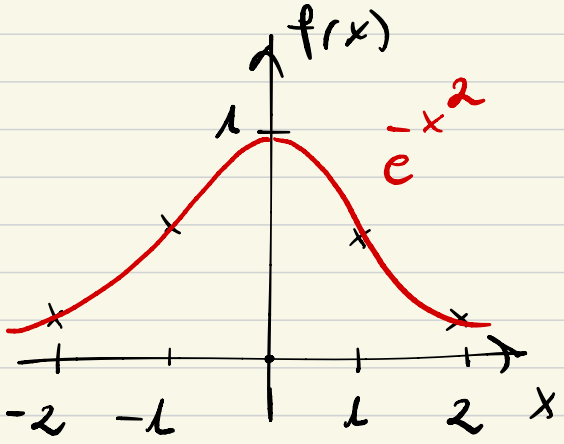
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

μ : mean value

σ : standard deviation

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

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



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

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

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

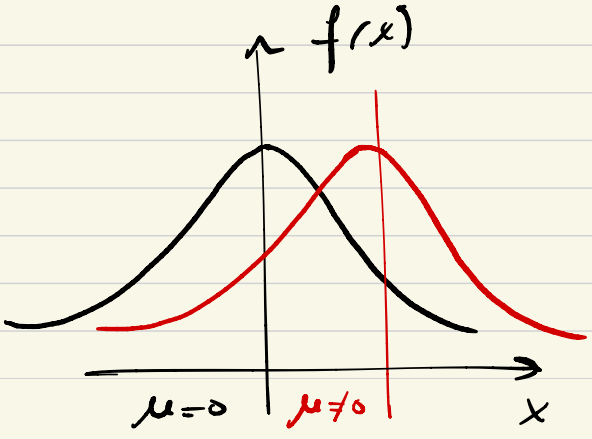
$x = \pm 2$; $f(\pm 2) = e^{-4} = \frac{1}{e^4} = 0.018$

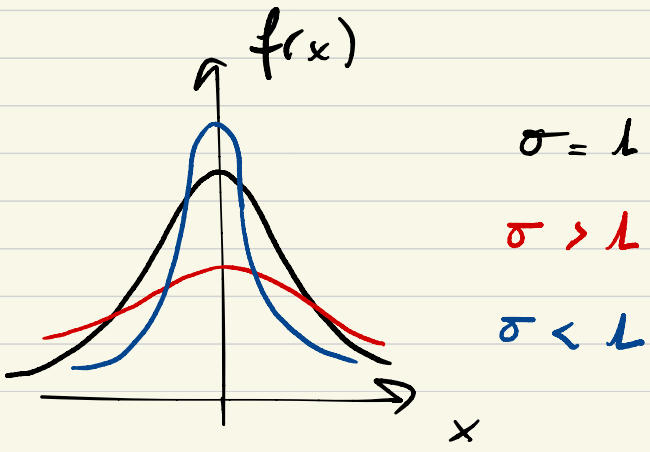
* Switch back on μ and σ

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

μ will displace the central value

σ will drive sharper/wider shape





* Add normalization factor

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

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

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

(*) Exercise

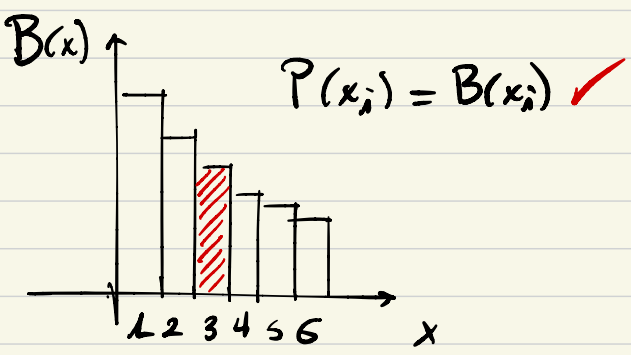
* for $f(x)$ to behave as a probability, we add a normalization factor

$$\left\{ f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \right\} \text{ Gaussian distribution; } \int_{-\infty}^{\infty} dx\, f(x) = 1$$

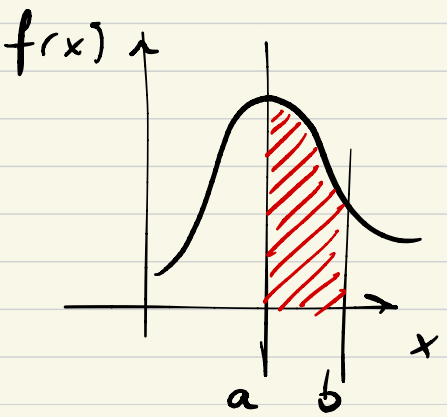
* Particular case $\mu = 0$; $\sigma = 1$

$$\left\{ f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2} \right\} \text{ Normal distribution } N(0,1)$$

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



$P(x_i) = B(x_i)$ ✓

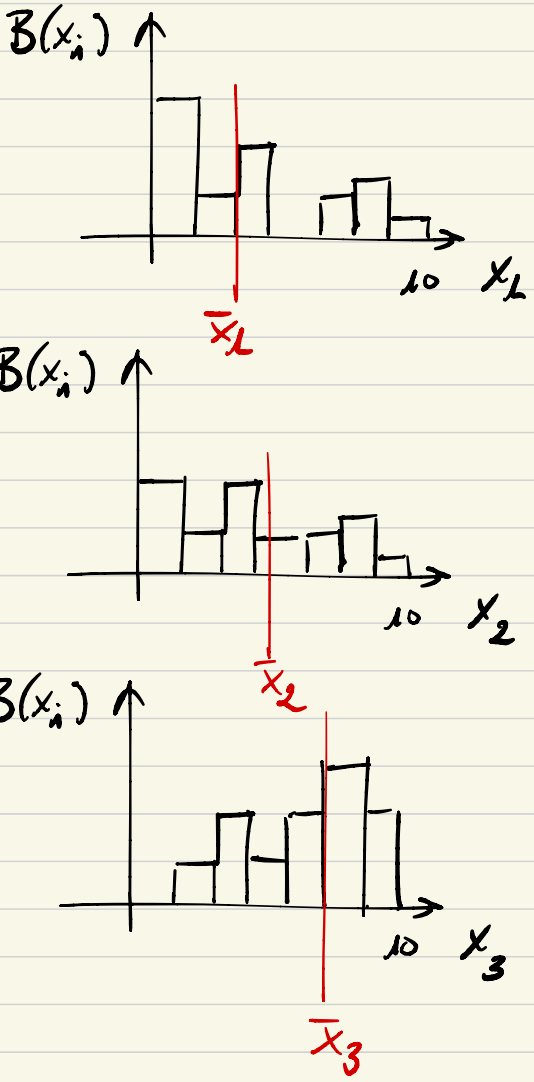


$P(a < x < b) = cdf(b) - cdf(a)$

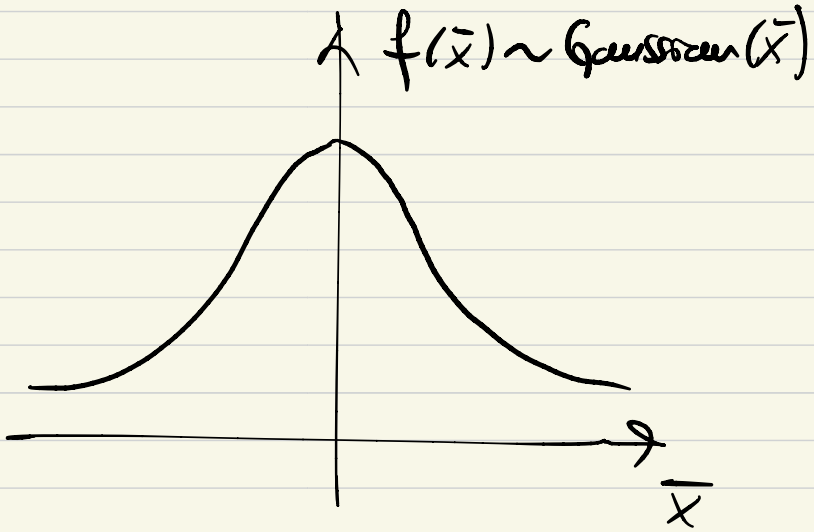
2.1 the Central Limit Theorem

"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 sample size n increases, the distribution of means $\bar{X}_i$ tends to a gaussian distribution



ii) Build new random variable

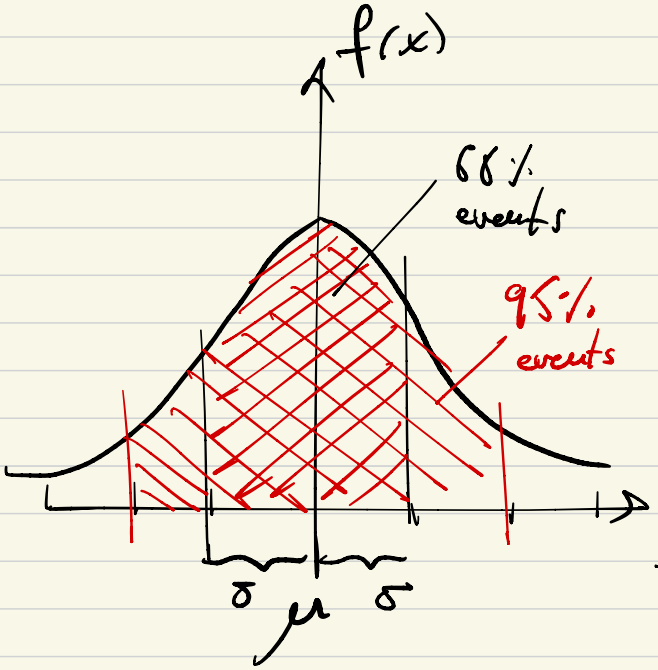
$$\bar{X} = \{\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_N\}$$

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

$$SE = \frac{\sigma_i}{\sqrt{n}}$$; as I increase n, $\downarrow$ 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: randomly distributed; some μ, σ.

standardized x: $z = \frac{x-\mu}{\sigma}$ 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

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

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.

iii) A bit of history

Karl Pearson (1857-1936). Pearson's "p" values for χ^2 test.

Mathematical statistics.

Ronald Fisher (1890-1962) "Statistical methods for research" (1925)

Example: Check if a die is biased.

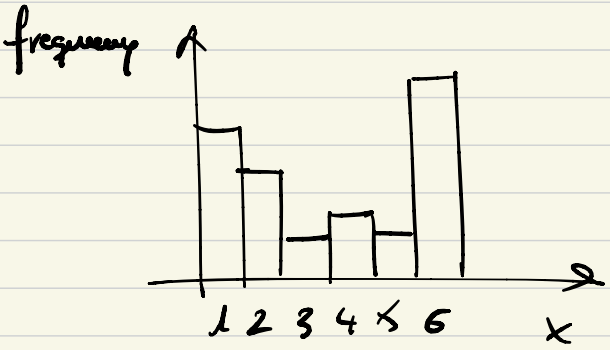
* You suspect your opponent is using a loaded die

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 $p = 1/6$

$B(43; 100, \frac{1}{6}) \sim 1$
cdf

Exercise

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

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

↳ p-value $< \alpha$. ∃ evidence to reject H_0.

Example: check if coin is fair

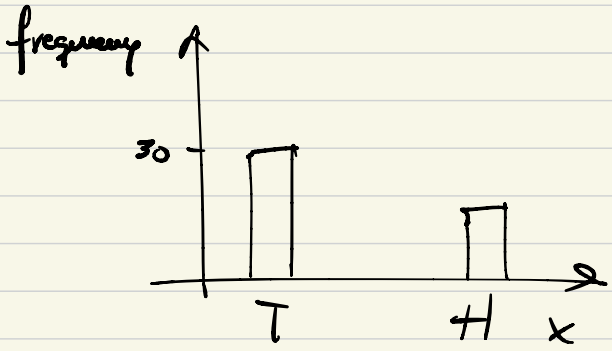
* You suspect your opponent is using a biased coin

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 $p = 1/6$

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

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.

(3.3) Comparing means. t-test

* Compare if two samples have significantly different means

Given two sets of observations x_1 and x_2, of sample size n_1 and n_2, and sample means $\bar{x}_1$ and $\bar{x}_2$.

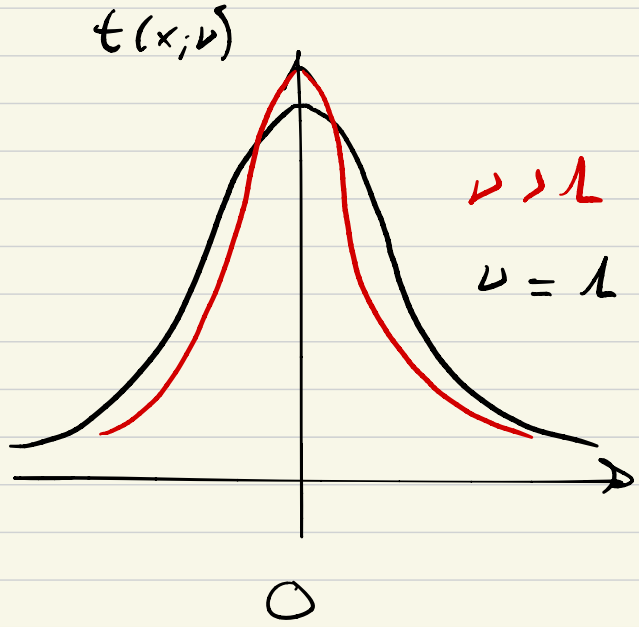
* t-test variable (t "statistic")

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

being s_p the pooled / sumed standard deviation

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

* Under null hypothesis of equal means, the t statistic follows a Student's t distribution with (n_1+n_2-2) dof.



For $\nu = 1$; $t(x;\nu)$ is refered to as the Cauchy distribution

$$f(x; x_0, \gamma) = \frac{1}{\pi} \frac{\gamma}{(x-x_0)^2 + \gamma^2}$$

* A bit of history

1876. t-distribution derived first by Helmert and Lüroth

1895. General form from Pearson, as Pearson type IV distribution

1908. William Sealy Gosset publishes under name "Student"

1925. Popularized by Fisher as "Student's distribution"

Example: Comparing means of 2 populations

* Check if the means of birth weights are significantly different between smokers and non smokers.

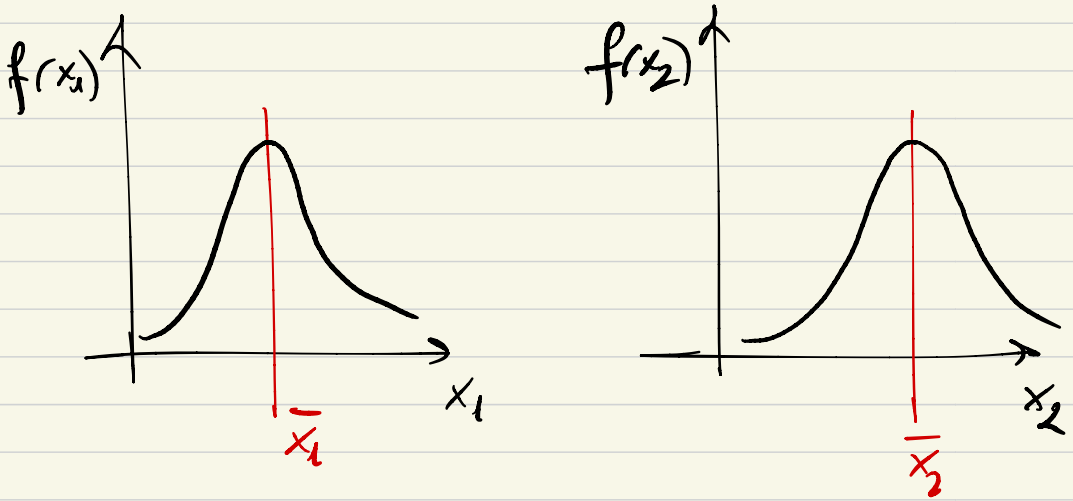
i) Formulate null hypothesis

H_0. $\nexists$ difference btw the means. $d = \bar{x}_1 - \bar{x}_2 = 0$

H_1. $\exists$ // . $d \neq 0$

Choose significance level α

ii) Collect data



n_1 non smokers — n_2 heavy smokers

Sample mean $\bar{x}_1$ — Sample mean $\bar{x}_2$

standard deviation s_1 — standard deviation s_2

iii) Compute t statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

iv) Compute d.o.f

$$\nu = n_1 + n_2 - 2$$

v) Check if t falls inside the 95% confidence interval.