Fundamentals of Probability

Introduction: what is probability?

- Branch of mathematics that deals with the likelihood or chance of events occurring
- Framework for reasoning about uncertainty and randomness
- Measure of how likely something is to happen, expressed as a number between 0 and 1
  - 0 → impossible, never happen
  - 1 → certain, definitely happen
- e.g. how likely is that the sun will rise tomorrow? how likely I'll get a 6 rolling a dice?

Do we really need probability in physics?

Several areas where this is useful:

- Handle measurement error/noise in experimental data
  - e.g. detector readouts in high-energy physics experiments
- Statistical mechanics: model collective instead of individual behavior
  - e.g. Maxwell-Boltzmann distribution for particle speed in gas
- Chaotic and complex systems
  - e.g. weather forecasts: range of possibilities varying initial conditions rather than deterministic prediction
- Quantum mechanics
  - e.g. inherent randomness of particles behavior, wave function, Heisenberg uncertainty principle
  - nature microscopic behavior, not just tool for measuring uncertainty!

Definition

We have multiple definitions of what probability is:

1. Axiomatic, Kolmogorov around 1933 – 3 axioms as general rules for computing probabilities
2. Classical (Combinatorial), Laplace around 18th century
3. Frequentist (empirical), von Mises around 20th century
4. Subjective, De Finetti around 20th century – degree of confidence an individual has in the occurrence of an event

---

Example: flipping a coin

The probability to get head in a coin flip according to different definitions is:

1. Classical (Combinatorial)

   $P(E) = \frac{Number\ of\ favorable\ outcomes}{Number\ of\ possible\ outcomes}$ → *two equally likely outcomes (head, tail), one is favorable: 1/2*

2. Frequentist (empirical)

   $P(E) = \lim_{n\to\infty} \frac{Number\ of\ times\ event\ E\ occurs}{Numb\quad of\ trials}$ → *flip a coin several times, count number of heads: 4/10*

3. Axiomatic, Kolmogorov around 1933
   3 axioms as general rules for computing probabilities
   → *sample space S = {H, T}; P(H) + P(T) = 1. Assuming a fair coin, then P(H) = P(T) = 1/2*

4. Subjective, De Finetti around 20th century
   degree of confidence an individual has in the occurrence of an event

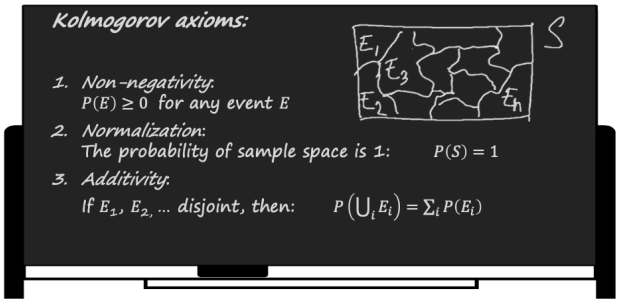
→ **what do you think the probability of heads is?**

1. Axiomatic definition

The axiomatic definition provides some rules to handle probability. From these we can derive further properties:

From these we can derive further properties:

- $P(\bar{A}) = 1 - P(A)$
- $P(A \cup \bar{A}) = 1$
- $P(\emptyset) = 0$
- $A \subset B \Rightarrow P(A) \leq P(B)$
- $\boldsymbol{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$



Axiomatic definition – limitations

Although elegant and useful, it poses a few practical and philosophical issues:

- Does not define the probability of individual events
- Requires a clearly defined sample space, not always the case; e.g. economics, human behavior, …
- Not applicable to non-measurable sets
- Build upon objective probability, not always practical; e.g. personal belief: P(«bus getting on time»)

EXERCISE: SHOW THAT

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

1. **Decomposition of $A \cup B$:**
The union of two sets can be split into disjoint subsets:
$A \cup B = A + B - (A \cap B)$,

where $A \cap B$ is the overlapping region that is counted twice when summing $P(A)$ and $P(B)$.

2. **Additivity of Probability:**
By the third axiom of probability, if two sets are disjoint (have no overlap), their probabilities can be summed:
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

2. Classical (combinatorial) definition

Classical definition turns out useful when:

- We can enumerate possible outcomes
- Outcomes are equiprobable

--> e.g. we have a bag with 3 blue balls and 2 red ones. What is the probability of drawing 2 blue balls at once?

Using axiomatic definition is impractical here, as we do not know the probability of drawing a ball of a given color.

- We can use the classic definition instead!

$$P(E) = \frac{\text{Number of favorable outcomes}}{\text{Number of possible outcomes}}$$

ELEMENTS OF COMBINATORICS

How to count and arrange n objects. Two key specs: order, repetitions

Permutations (order important)

- w/o repetition:

$$n!$$

Example: How many ways can you arrange 3 letters A, B, C?
$3! = 3 \times 2 \times 1 = 6 \quad (\text{ABC},\ \text{ACB},\ \text{BAC},\ \text{BCA},\ \text{CAB},\ \text{CBA})$

- w/ repetition:

$$\frac{n!}{k_1!k_2!\dots k_r!}$$

Example: How many ways can you arrange A, A, B?
$\frac{3!}{2! \cdot 1!} = \frac{6}{2} = 3 \quad (\text{AAB},\ \text{ABA},\ \text{BAA})$

Dispositions (order important)

- w/o repetition:

$$\frac{n!}{(n-k)!}$$

Example: How many ways can you pick and arrange 2 items from A, B, C?
$\frac{3!}{(3-2)!} = \frac{6}{1} = 6 \quad (\text{AB},\ \text{AC},\ \text{BA},\ \text{BC},\ \text{CA},\ \text{CB})$

- w/ repetition:

$$n^k$$

> Example: How many ways can you pick and arrange 2 items from A, B, C with repetition?
> $3^2 = 9 \quad (\text{AA},\ \text{AB},\ \text{AC},\ \text{BA},\ \text{BB},\ \text{BC},\ \text{CA},\ \text{CB},\ \text{CC})$

Combinations (order not important)

- w/o repetition:

$$C_{n,k} = \frac{n!}{k!(n-k)!}$$

> Example: How many ways can you choose 2 items from A, B, C?
> $C_{3,2} = \frac{3!}{2! \cdot (3-2)!} = \frac{6}{2 \cdot 1} = 3 \quad (\text{AB},\ \text{AC},\ \text{BC})$

- w/ repetition:

$$C'_{n,k} = \frac{(n+k-1)!}{k!(n-1)!}$$

> *Example: How many ways can you choose 2 items from A, B, C with repetition?*
> $C'3, 2 = \frac{(3+2-1)!}{2! \cdot (3-1)!} = \frac{4!}{2! \cdot 2!} = \frac{24}{4} = 6 \quad (\text{AA},\ \text{AB},\ \text{AC},\ \text{BB},\ \text{BC},\ \text{CC})$

Back to classical definition.. and the question:

- e.g. we have a bag with 3 blue balls and 2 red ones. What is the probability of drawing 2 blue balls at once?

Using the axiomatic definition is impractical here, as we do not know the probability of drawing a ball of a given color.

We can use the classic definition instead! Assuming each ball is equiprobable, we just need:

- How many ways of extracting 2 balls out of 5?
- How many of them contain 2 blue balls?

$$P(\text{"2 blue balls"}) = \frac{C_{3,2}}{C_{5,2}} = \frac{\binom{3}{2}}{\binom{5}{2}} = \frac{3!}{2!1!} \cdot \frac{5!}{2!3!} = \frac{3}{10}$$

Classic definition – limitations

Although intuitive, it poses a few practical and philosophical issues:

- Tautological, self-referential: what does it mean by "equiprobable"?
- What if events are not equally likely? For example, (P(\text{"gold medal at Olympics final"}))
- Enumerating outcomes is not always feasible or even possible! (e.g. infinite sets)

- Does not apply to empirical data (e.g. (P(\text{"taller than 1.80m"}))

3. Frequentist definition

Frequentist definition $P(E) = \lim_{n\to\infty} \frac{Number\ of\ times\ event\ E\ occurs}{Number\ of\ trials}$

Frequentist definition assumes:

- Objective probabilities exist
- We can collect data about a phenomenon and count how many times an event of interest is observed
- As the number of trials (data) grows, the relative frequency approaches the true probability of the event
- Experiments can be repeated under identical conditions
  - e.g. What is the probability of an atom decaying in the next year?
    - Take many atoms and put them under same initial conditions
    - Observe them for a year
    - Count how many of them have decayed
  - As the number of atoms grows, we have:

$$\lim_{n\to\infty} P(\text{decay in 1 year})$$

$$P(E) = \lim_{n\to\infty} \frac{\text{Number of times event E occurs}}{\text{Number of trials}}$$

Frequentist definition limitations

• Experiments must be repeated under identical conditions, not always possible; e.g. difficult to control external factors
• Requires large number of trial for accurate approximation
• Consequently bad-suited for rare phenomenon, especially one-off events

4. Subjective definition

This definition is based on a quantification of the degree of belief. For this, we use a fair bet:

«P(A) = fraction of payout, Y, one would bet on A in order not to neither win nor lose money»

- Based on personal belief and knowledge
- Useful for unique, non-repeatable events; e.g. political election, betting, …
- Only option in many practical situations
- e.g. What is the probability that tomorrow will rain?
  - Weather forecasts report good weather
  - Today is sunny and warm
  - I have outdoor activities planned for tomorrow

-- Pessimistic: It would be fair to bet 10€ to win 100€ as, for me, rain is 9 times more likely: P(A) = 10/100
-- Optimistic: It would be fair to bet 99€ to win 100€ as, for me, sun is 99 times more likely: P(A) = 99/100

Subjective definition limitations

- Still hard to quantify, opinions change
- Naturally variable, personal

Conditional probability & independence

Two fundamental concepts when operating on probabilities are conditional probability and independence:

- **Conditional probability** is the probability of an event occurring based on a given prior knowledge. I.E. $P(A|B)$ is the probability of A happening given that we know B has already happened.
    - Conditional probability is defined as: $P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) \neq 0$
    - **Example:** Rolling a number lower than 3 given that the outcome is even:

$$P(n < 3 | n \text{ even}) = \frac{P(n < 3 \cap n \text{ even})}{P(n \text{ even})} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

- **Independence:** Two events are independent if knowing something about one tells us nothing about the other.
    - Mathematically:

$$A, B \text{ independent} \implies P(A \cap B) = P(A)P(B)$$

    - Note that if A, B are independent, $P(A|B) = P(A)$
    - Important: independent disjoint $P(A \cap B) = \emptyset$

Frequentist VS Subjective probability and Bayes' Theorem

Outline

- Bayes' Theorem
- Frequentist VS Bayesian statistics
- Examples

How to update probability based on new evidence?

Bayes' theorem

Bayes' theorem provides a nice mechanism to update probability in light of new evidence:

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

- **Prior Probability, P(A)**: Initial belief before seeing evidence
- **Marginal Likelihood, P(B)**: Overall probability of the evidence
- **Likelihood, $P(B|A)$**: Probability of evidence given the hypothesis
- **Posterior Probability, P(A|B)**: Updated probability after observing evidence

Key Insights:

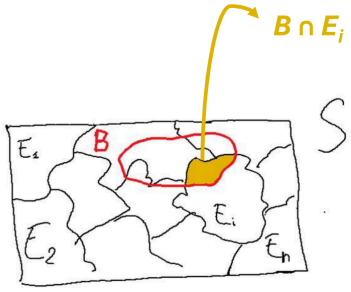
- Bayes' Theorem updates beliefs based on new evidence
  → resembles how we think
- It accounts for both the strength of the evidence and prior knowledge
- Note: prior belief/knowledge is not necessarily subjective probability

Law of total probability

Express the probability of an event B in terms of a disjoint partition of the sample space S:

- Partition the sample space S into disjoint subsets E_i so that: $\cup E_i = S$
- Then a subset B of S can be expressed as:

$$B = B \cap S = B \cap \left(\bigcup E_i\right) = \bigcup (B \cap E_i)$$



- Leveraging conditional probability, we can re-write:

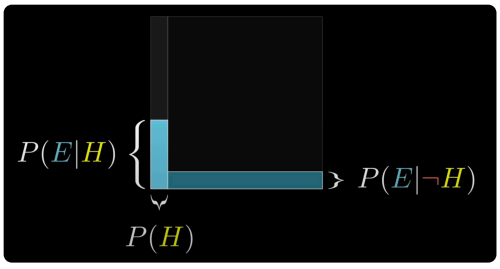
$$\boxed{\mathrm{P(B)}} = \mathrm{P}\big(U_i(B \cap E_i)\big) = \Sigma_i P(B \cap E_i) = \boxed{\Sigma_i P(B|E_i)P(E_i)} \quad \textbf{Law of total probability}$$

> Here we used the definition of conditional probability: is a measure of the probability of an event occurring, given that another event (by assumption, presumption, assertion or evidence) is already known to have occurred
>
> - **Scenario**: A bag contains 10 marbles: 6 red, 3 blue, 1 green.
> - **Events**:
>   - A: Drawing a red marble.
>   - B: Drawing a marble that is not green.
>
> **Steps**:
>
> 1. **Calculate** $P(B)$:
>   Marbles not green = 6 red + 3 blue = 9.
>
> $$P(B) = \frac{\text{Marbles not green}}{\text{Total marbles}} = \frac{9}{10}$$
>
> 2. **Calculate** $P(A \cap B)$:
> $P(A \cap B)$ = Probability of drawing a red marble (all red marbles are not green).
>
> $$P(A \cap B) = \frac{\text{Red marbles}}{\text{Total marbles}} = \frac{6}{10} = \frac{3}{5}$$
>
> 3. **Calculate** $P(A \mid B)$:
>   Using the formula for conditional probability:
>
> $$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{3}{5}}{\frac{9}{10}} = \frac{2}{3}$$

Thus, Bayes' theorem becomes:

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{\Sigma_i P(B|E_i)P(E_i)}} \quad \textbf{Bayes' THM using law of total probability}$$

This is a graphical rapresentation



- E is the evidence
- H is the hypothesis

Bayes' theorem: example

Suppose you want to know the probability of having a disease (A) given that you tested positive (B) for it:

- **Prior Probability** $(P(A) = P(\text{disease}))$: reflects our belief in the hypothesis before seeing any evidence, e.g., probability of the disease in the population, independent of the test
- **Likelihood** $(P(B|A) = P(\text{positive test}|\text{disease}))$: how probable the evidence is, assuming the hypothesis is true, i.e., the probability of testing positive, given that you actually have the disease
- **Marginal Likelihood** $(P(B) = P(\text{positive test}))$: probability of the evidence, i.e., overall probability of testing positive, including both correct and incorrect outcomes
  → law of total probability:

$$P(B) = P(B|A) \cdot P(A) + P(B|\neg A) \cdot P(\neg A)$$

- **Posterior Probability** $(P(A|B) = P(\text{disease}|\text{positive test}))$: updated probability after observing the evidence, i.e., probability of having the disease given the test result

In practice:

- Suppose 1% of the population has the disease → $P(\text{disease}) = 0.01$
- Suppose the test has:
  - 90% sensitivity, i.e., it correctly identifies 90% of diseased testers → $P(\text{positive test}|\text{disease}) = 0.90$
  - 5% false positive rate, i.e., 5% of healthy people test positive → $P(\text{positive test}|\neg\text{disease}) = 0.05$

We can compute the posterior probability using Bayes' Theorem:

$$P(\text{disease}|\text{positive}) = \frac{P(\text{positive}|\text{disease})P(\text{disease})}{P(\text{positive})}$$

Where $P(\text{positive})$ is the marginal likelihood:

$$\begin{aligned} P(\text{positive}) &= P(\text{positive}|\text{disease})P(\text{disease}) + P(\text{positive}|\neg\text{disease})P(\neg\text{disease}) \\ &= 0.9 \cdot 0.01 + 0.05 \cdot 0.99 = 0.0585 \end{aligned}$$

Therefore:

$$P(\text{disease}|\text{positive}) = \frac{0.9 \cdot 0.01}{0.0585} = 0.1538 \approx 15\%$$

Exercise

A beam of particles consists of a fraction 10^{-4} electrons and the rest photons. The particles pass through a double-layered detector which gives signals in either zero, one or both layers. The probabilities of these outcomes for electrons (e) and photons (γ) are

$$\begin{array}{ll} P(0\,|\,\mathrm{e}) = 0.001 & P(0\,|\,\gamma) = 0.99899 \\ P(1\,|\,\mathrm{e}) = 0.01 & P(1\,|\,\gamma) = 0.001 \\ P(2\,|\,\mathrm{e}) = 0.989 & P(2\,|\,\gamma) = 10^{-5}. \end{array} \quad \text{and}$$

(a) What is the probability for the particle to be a photon given a detected signal in one layer only?

(b) What is the probability for a particle to be an electron given a detected signal in both layers?

What probability interpretation should we use?

We have seen several definitions. Which one should we use?

- **Axiomatic definition**
  - Rules to handle probability mathematically
  - Always applies but not practical as we do not have $P(E_i)$
- **Classical**
  - Useful when we can enumerate favorable and total outcomes
  - Outcomes are equiprobable
  - Not practical (or even unfeasible) with large/infinite sample spaces
- **Frequentist**
  - Useful when we can perform repeated experiments under the same conditions
  - Note: the more trials, the better the approximation!
- **Subjective**
  - Allows dealing with one-off or rare events
  - Based on personal belief and knowledge, therefore questionable
    → **Bayes' theorem makes it more rigorous and mathematically grounded**

What is statistics?

Statistics is the science of collecting, analyzing, interpreting, presenting, and organizing data.

- Provides tools and methodologies to make sense of raw data and draw conclusions
  → turns data into meaningful information

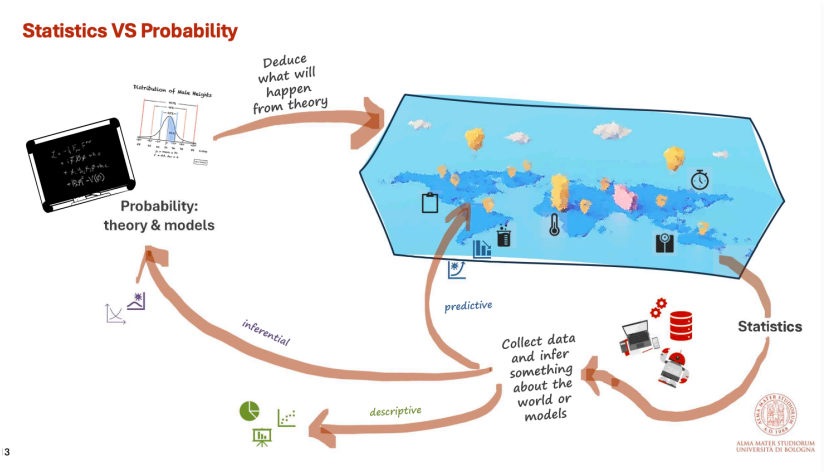
Key areas:

- **Descriptive Statistics:** summary and description of main features in data, e.g., mean, median, variance, standard deviation, coefficient of variation

- **Inferential Statistics:** goes beyond the data to make predictions or inferences about a larger population
  → hypothesis testing, confidence intervals, and predictions based on observed data
- **Prediction:** W hat will happen given my model and the given data?
- **Inference:** What can I learn from the data about my model?

Statistics helps us answer questions like:

- Is a new drug effective in treating a disease?
- What is the average strength of a material under stress?
- Is the observed signal a new particle, or due to random background fluctuations?



Frequentist statistics

Frequentist statistics is a framework that:

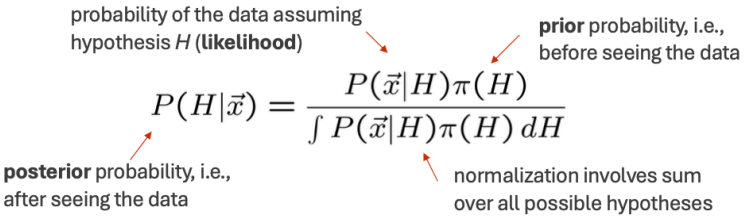
- Builds upon the frequentist definition of probability: long-run frequencies of events
  → repeatable experiments
- The concept of probability is strictly tied to data:
  - We cannot answer: "P(Higgs boson exists)?"
    - $P(\text{Higgs boson exists})$ is either 0 or 1, we do not know which
  - We cannot answer: "P(\text{Higgs boson exists} | \text{data})?" either
  - What we can do: "P(\text{data} | \text{Higgs boson exists})?"
- Accepted theories/models are those most compatible with experimental data.

Bayesian statistics

Bayesian statistics is a framework that:

- Builds upon the subjective definition of probability: degree of belief about a hypothesis

- Exploits Bayes' theorem to update our knowledge based on data:



  → Actually measuring $P(\text{Higgs boson exists}|\text{data})$!
- However, the prior choice is subjective.

Frequentist VS Bayesian statistics

In summary, two alternative interpretations:

- Frequentist framework works on the likelihood of data given hypotheses: $P(\text{data}|H)$
- Bayesian approach works on prior update given data: $P(H|\text{data})$
- Frequentists take decisions based on the likelihood.
- Bayesians take decisions based on the posterior.

Random variables

Outline

- Random variables
- Examples of univariate distributions
- Moments and characteristic function

Random variables

A random variable is a mathematical object that maps a numerical value to each outcome of a random process:

- A R.V. X is composed of a probability triplet:
  - Sample space, S: set of all possible outcomes
  - Event space, E: set of all events, i.e., all subsets of S
  - Measurable function, P: maps each event to its probability
- Example:
  - R.V. X represents the random process of rolling dice
  - Sample space: $S = \{1, 2, 3, 4, 5, 6\}$
  - Event space: $E = \{\{1\}, \{2\}, \ldots, \{1, 2\}, \{1, 3\}, \ldots\}$
  - Probability function, $P : F \to \mathbb{R}$ that associates to each event its probability $\to$ distribution

Notation

> **Notation**
> Capital latin letters are used to denote random variables, e.g. X is R.V. for coin flip
> Their realizations are denoted with the corresponding lowercase, e.g. x_1="tails", x_2="heads"

Two types of R.V.s depending on the sample space:

Discrete

- S is finitely or infinitely **countable**
- P is called probability **mass** function, ***p*** (pmf):
  - $P(X = x_i) = p_i$
  - $\sum_{x_i \in S} P(X = x_i) = 1$
- **Cumulative distribution** function, ***F*** (also «funzione di ripartizione»):

$$P(X \leq x) = \sum_{x_i \leq x} p(x_i) \equiv F(x)$$

Continuous

- S is **uncountable**
- P is called probability **density** function, ***f*** (pdf):
  - $\boldsymbol{P(X = x_i) = 0}$ **!!**
  - $P(X \in [x, x + dx]) = \int_x^{x+dx} f(x)\, dx$
  - $\int_S f(x)\, dx = 1$
- **Cumulative distribution** function, ***F*** (cdf):

$$P(X \leq x) = \int_{-\infty}^{x} f(z)\, dz \equiv F(x)$$

- Alternatively define pdf as:

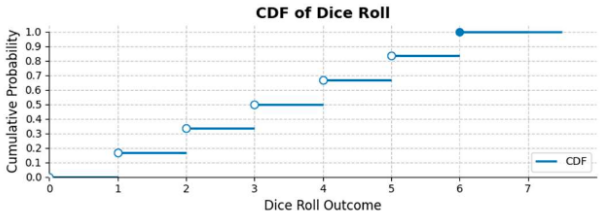
$$f(x) = \frac{dF(x)}{dx}$$

Cumulative distribution properties

The cumulative distribution has several important properties:

- Non-decreasing
- Right-continuous
  - Step function for discrete R.V.s
  - May be a step function also in the continuous case

$$\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1$$

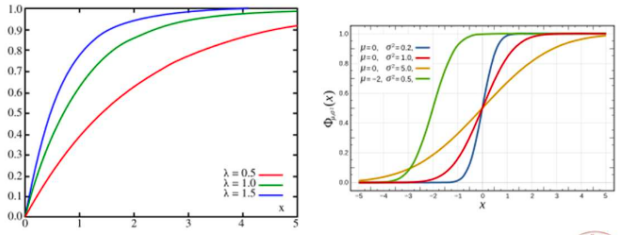


For continuous R.V.s, given constants a, b such that $a < b$:

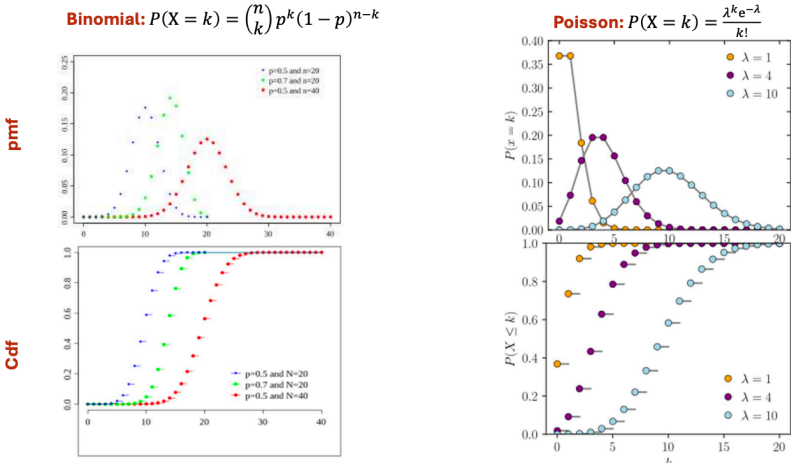
$$F(b) - F(a) = P(a < X \leq b) = \int_a^b f(x)dx$$

Note: in practice, we can use $<$ and $\leq$ indistinguishably, as adding a point to the integral does not affect the result.

A similar result holds for discrete R.V.s, but more attention needs to be paid to inequalities.

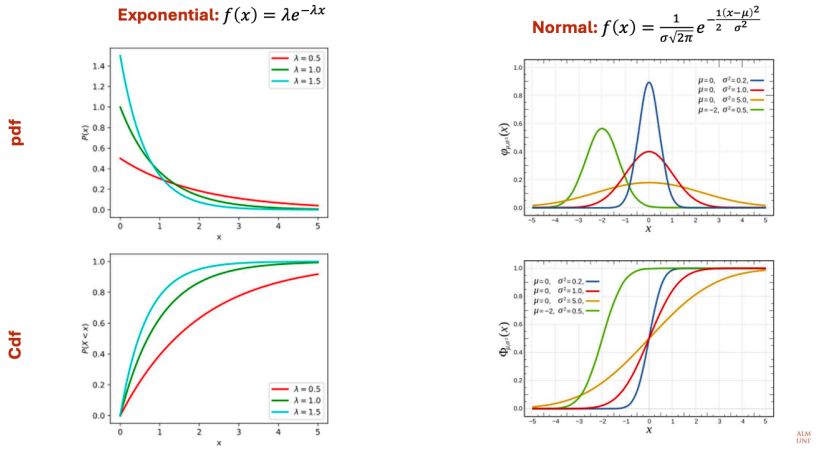


Examples of DISCRETE random variable distributions



.

Examples of CONTINUOUS random variable distributions



How to summarize a distribution?

- Distributions enclose all information about a random variable.
- However, sometimes **we do not need all that information.**
- Can we retrieve some synthetic measure of relevant features?
  - Central tendency
  - Spread
  - Symmetry
  - Behavior in the tails
  - Also useful for quantitative comparisons

Summary statistics: Central tendency and spread

Two useful properties often used are:

- **Expected value**, μ:
  - Discrete:

$$E[X] = \sum x \cdot p(x)$$

  - Continuous:

$$E[x] = \int_{-\infty}^{\infty} x f(x) dx$$

  - E is a linear operator called **expectation.**
    → weighted sum (or integral), with probability (or probability density) as weight
    → measures central tendency of the distribution

- **Variance**, σ^2:
  - Discrete:

$$V[X] = \sum (x-\mu)^2 \cdot p(x)$$

if $\mu = \mathbb{E}(X)$

$$V[X] = \sum (x - \mathbb{E}(x))^2 p(x)$$

  - Continuous:

$$V[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

  - **Variance** measures spread around the expectation.

Expectation operator properties

The expectation is a linear operator, which implies several properties:

- $E[aX] = aE[X]$ where a is a constant value.
- $E[X+Y] = E[X] + E[Y]$
  - This is true irrespectively of whether X, Y are independent.
- If $X \perp Y$, then $E[XY] = E[X]E[Y]$
  - However, $E[XY] = E[X]E[Y] \nRightarrow X \perp Y$
- $V[X] = E[X^2] - (E[X])^2$

Proof:

$$\begin{aligned} V(X) &= E\left[(X - E(X))^2\right] \\ &= E[X^2 - 2XE(X) + E(X)^2] \\ &= E[X^2] - 2E(X)E(X) + E(X)^2] \\ &= E(X^2) - E(X)^2 \end{aligned}$$

with the red highlight being constant, so replaceable with μ.

Moments of a distribution:

In general, we can look at moments as summary statistics. For a continuous R.V. X, the moment of order m about c is defined as:

$$\mu_m = E[(X-c)^m] \equiv \int_{-\infty}^{\infty} (x-c)^m f(x) dx$$

where c is a constant value.

- For a discrete R.V. X, just replace the integral with a sum and the pdf with a pmf.

.

.

.

.

.

IMPORTANT MOMENTS

- **Raw (initial) moments** → $c = 0$:

$$\mu_m = E[X^m]$$

  - The order 1 (m = 1) moment is the expected value:

$$\mu = E[X]$$

  → central tendency of a distribution

- **Central moments** → $c = \mu$:

$$\mu_m = E[(X - \mu)^m]$$

  - The order 2 (m = 2) moment is the variance:

$$\sigma^2 = V[X]$$

  → spread around μ

- **Standardized moments** → $c = \mu$:

$$\mu_m = \frac{E[(X - \mu)^m]}{\sigma^m}$$

  - The order 3 (m = 3) moment is the skewness → measures lopsidedness.
  - The order 4 (m = 4) moment is the kurtosis → measures tail heaviness.

Exercise

Consider a continuous random variable X and two constants, α, β. Starting from the definition of expected value, show that:

- $E[\alpha X + \beta] = \alpha E[X] + \beta$
- $V[\alpha X + \beta] = \alpha^2 V[X]$

Characteristic functions

The characteristic function $\phi_X(k)$ of an r.v. X is defined as the expectation value of e^{ikX} (similar to the Fourier transform of X):

$$\phi_X(k) = E[e^{ikX}] = \int_{-\infty}^{\infty} e^{ikx} f(x) dx$$

- Useful for finding moments and deriving properties of sums of R.V.s.
- For well-behaved cases (true in practice), the characteristic function is equivalent to the pdf and vice versa, i.e., given one, you can, in principle, find the other.

Moments from the characteristic function

Given a random variable Z, we can derive the moments from its characteristic function. To find the m-th moment:

- Differentiate m times $\phi_z(k)$
- Evaluate at k = 0

$$\begin{aligned}\left.\frac{d^m}{dk^m}\phi_z(k)\right|_{k=0} &= \left.\frac{d^m}{dk^m}\int e^{ikz} f(z)dz\right|_{k=0} \\ &= i^m \int z^m f(z)dz \\ &= i^m \mu_m\end{aligned}$$

where $\mu_m = E[X^m]$ is the m-th initial moment of z.

→ From the characteristic function, we can derive moments even without an explicit formula for the pdf.

Exercise

- X and Y are two independent random variables, and $Z = X + Y$ is derived as their sum. Prove that the characteristic function $\phi_Z(k)$ is the product $\phi_Z(k) = \phi_X(k)\phi_Y(k)$.
- Does this hold for a general linear combination of independent random variables?
  i.e., if $Z = \sum a_j X_j$, then $\phi_Z(k) = \prod \phi_{X_j}(a_j k)$.

Popular discrete and continuous distributions

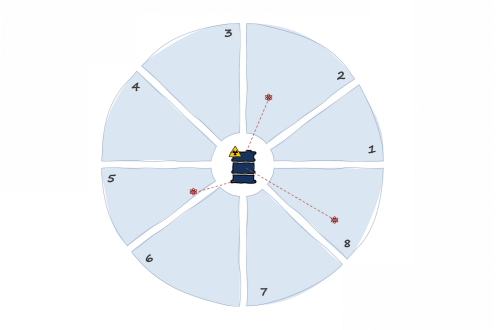
Outline

- How can we use R.V.s in practice?
- Examples of R.V.s: Uniform, Bernoulli, Binomial, Poisson, Exponential, Gaussian, Student's t, chi square
- More examples of R.V.s: Beta, Gamma, Breit-Wigner (Cauchy), Landau

How do we use R.V.s for modelling real processes?

1. Uniform distribution: discrete

Example: detecting alpha particles emitted from a radioactive sample, assuming isotropic emissions



- We divide the space into 8 equal regions, each covered by a detector
  → in which region will we observe the next emitted particle?
- Cannot tell in advance: random process!
- Each region has the same probability since isotropic

Mathematical formulation:

- X is R.V. describing the region where the alpha particle is emitted
- Sample space, S: {1, 2, 3, 4, 5, 6, 7, 8}
  → **countable**, so X is a discrete R.V.

- **pmf**, $p(X = x) = \frac{1}{8}$ for all $x \in S$
- **cdf**, $F(X \leq x)$?
  - $\lim_{x \to -\infty} F(x) = 0$
  - $\lim_{x \to \infty} F(x) = 1$
  - $F(x) = \frac{x}{8}$, $x \in S$

Expected value:

$$E(X) = 1 \cdot \frac{1}{8} + 2 \cdot \frac{1}{8} + \cdots + 8 \cdot \frac{1}{8} = 4.5$$

Variance:

$$V(X) = \sum (x - E(X))^2 \cdot p(x) = 5.25$$

- Is there easier way? Yes:

$$V[X] = E[X^2] - (E(X))^2$$

In general, the **discrete Uniform** distribution describes random processes with a **finite** number of outcomes, all equiprobable (e.g., fair die roll, randomly pick one out of n elements)

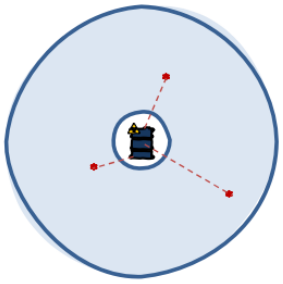
So $X \sim \text{Uniform}(a, b)$ and the sample space, S: $[a, b] = \{a, a+1, a+2, \ldots, a+n-1\}$
→ n elements from a to b, spaced by 1 unit

Then

- pmf, $p(X = x) = \frac{1}{n}$ for all $x \in S$
- cdf, $F(X \leq x)$:
  - 0, $x < a$
  - $\frac{x-a+1}{b-a+1}$ for $a \leq x \leq b$
  - 1, $x > b$

- Expected value: $E(X) = \frac{a+b}{2}$
- Variance: $V(X) = \frac{(b-a+1)^2-1}{12}$
- Characteristic function: $\varphi(k) = \frac{e^{ika}(1-e^{ik(b-a+1)})}{(b-a+1)(1-e^{ik})}$

1. Uniform distribution: continuous



Now imagine we have a single detector covering all the space around the sample

We have that $X \sim \text{Uniform}(a, b)$ and the sample space, S: $[a, b]$
→ this time is the continuous interval! e.g. $[0, 2\pi]$

- pdf, $f(X = x) = \frac{1}{b-a}$ for $a \leq x \leq b$; 0 otherwise
- cdf, $F(X \leq x)$:
  - 0, $x < a$
  - $\frac{x-a}{b-a}$ for $a \leq x \leq b$
  - 1, $x > b$

- Expected value: $E(X) = \frac{a+b}{2}$
- Variance: $V(X) = \frac{(b-a)^2}{12}$

- Characteristic function: $\varphi(k) = \frac{e^{ika}(1-e^{ik(b-a)})}{(b-a)(1-e^{ik})}$

---

2. Bernoulli distribution

Example: a single coin flip

- Only two possible outcomes: heads or tails
  → denoted by 1, "success", and 0, "failure", respectively
- We denote by p the probability of success, e.g., $p = 0.3$.

Mathematical formulation:
X is R.V. describing the outcome of a coin flip, $X \sim \text{Bern}(p)$ with Sample space, S: $\{0, 1\}$

- pmf, $p(X = x)$:
  - p for $x = 1$
  - $(1 - p) = q$ for $x = 0$
- cdf, $F(X \leq x)$:
  - 0, $x < 0$
  - $1 - p$ for $0 \leq x \leq 1$
  - 1, $x > 1$
- Expected value: $E(X) = p \cdot 1 + (1 - p) \cdot 0 = p$
- Variance: $V(X) = p - p^2 = p(1 - p)$
- Characteristic function: $\varphi(k) = 1 - p + pe^{ik}$

When to use: Use the Bernoulli distribution for experiments with only two possible outcomes (e.g., success/failure) where each trial is independent and has a fixed probability of success.

2.1 Binomial distribution

Example: n coin flips

- n independent trials with binary outcomes: heads or tails
- Each flip has the same probability of success, p

Mathematical formulation:
X is R.V. describing the number of successes in n independent coin flips,
$X \sim \text{Bin}(n, p)$ with Sample space, S: $\{0, 1, \ldots, n\}$
→ countable, so X is a discrete R.V.

- pmf $p(X = k)$?
  - Note: $X = \sum Y_i$, where $Y_i \sim \text{Bern}(p)$
  - Recall that, if A ⊥ B then: $P(A \cap B) = P(A)P(B)$

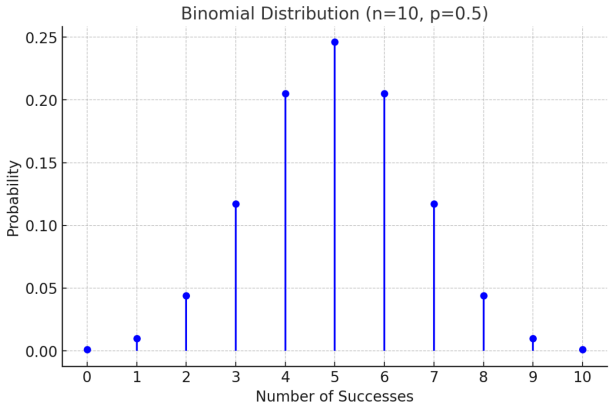
- Recall that, if A, B disjoint: $P(A \cup B) = P(A) + P(B)$
- Can we derive $P(X = 0)$? And $P(X = 1)$? And $P(X = 2)$? And $P(X = k)$

So we end up with

$$p(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

which makes sense because

1. The number of ways to arrange k successes in n trials is the binomial coefficient
2. The probability in n trials has to be multiplied for each k success and $(n-k)$ not success



- **cdf**, regularized incomplete beta function
- Expexted value: $E[X] = np$
- Variance: $V[X] = np(1-p)$
- Characteristic Function: $\phi_X(k) = [1 + p(e^{ik} - 1)]^n$

When to use: Use the Binomial distribution for a series of independent trials with two outcomes each (e.g., success/failure), where the probability of success is constant across trials, and you want to model the number of successes out of a fixed number of trials n.

---

3. Poisson distribution

Example: Studying the decay of a radioactive isotope with a known average decay rate.

- We have a sample of radioactive material.
  - **Question**: How many decays will occur in the next minute?
- The process is random, so we cannot tell in advance.
- However, we know the average decay rate, denoted by λ (e.g., 5 decays per minute).

Mathematical formulation:

- X is an R.V. representing the number of decays in the next minute, $X \sim \text{Poisson}(\lambda)$.
- Sample space, $S = \{0, 1, 2, 3, \ldots\}$
  - This is infinitely countable, so X is a discrete R.V.
- **pmf:** $$
  p(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{for } k \in S

  ![[Pasted image 20241228160810.png|400]]
- **Moments:**
  - **Expected value:** $E(X) = \lambda$.
  - **Variance:** $V(X) = \lambda$
- **cdf:** Calculated using the regularized gamma function.
- **Characteristic function:**

$$\varphi(t) = e^{\lambda(e^{it}-1)}$$

When to use: The Poisson distribution is suitable for modeling random counting processes where rare events occur at a known average rate (e.g., phone calls in an hour or accidents per day).

- If $X \sim \text{Poisson}(\lambda)$, then:
  - λ is the average count in the time interval, also called the **intensity**.
  - **Example:** Expressing λ as:

$$\lambda = r \cdot t$$

    where r is the event rate, and t is the duration of the time interval. Here, λ remains constant over time.
- Sample space, $S = \mathbb{N} = \{0, 1, 2, \ldots\}$

Relationship between Poisson and Binomial distributions

> Counting resembles repeated observations of whether an event has occurred (success) or not (failure).

- **Poisson:** models counting processes for rare events.
- **Binomial:** counts the number of successes in a series of independent trials.

To understand their relationship, consider:

- The Binomial distribution counts the number of successes in independent repeated trials:
  - p: probability of success in a single trial.
  - n: number of trials.

Hint: Try comparing the characteristic functions:

1. **Binomial characteristic function:**

$$\varphi_{\text{Bin}}(k) = \left[1 + p(e^{ik} - 1)\right]^n$$

2. **Poisson characteristic function:**

$$\varphi_{\text{Poisson}}(k) = e^{\lambda(e^{ik}-1)}$$

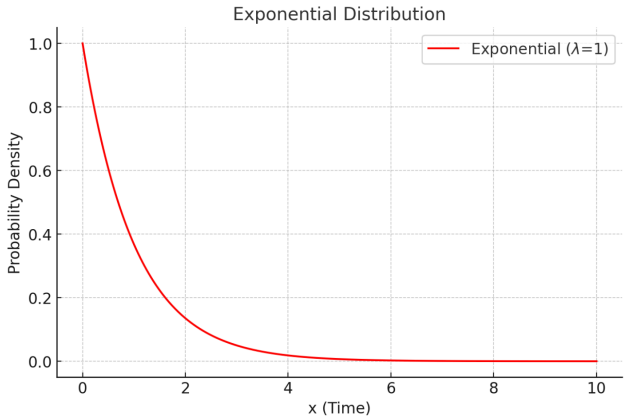
   By letting $p = \frac{\lambda}{n}$ and considering the limit as $n \to \infty$, the Binomial distribution approaches the Poisson distribution.

> In general the Binomial converges to the Poisson for $n \to \infty$ and $p \to 0$

---

4. Exponential Distribution

- **Example:** Studying the decay of a radioactive isotope, known average rate.
  - What is the waiting time between successive decays?
  - We know: 5 decays per minute, on average (λ).
- **Mathematical formulation:**
  - X is R.V. for the waiting time between successive decays, $X \sim Exp(\lambda)$.
  - Sample space, S: $[0, \infty)$.
    - $\Rightarrow$ uncountable, so X continuous R.V.
  - **pdf,** $p(X = x) = \lambda e^{-\lambda x}$, $x \geq 0$
  - **cdf,** $F(X \leq x) = 1 - e^{-\lambda x}$, $x \geq 0$



  - **Characteristic function:** $\phi(t) = \frac{\lambda}{\lambda - it}$.
  - **Expected value:** $E(X) = \frac{1}{\lambda}$
  - **Variance:** $V(X) = \frac{1}{\lambda^2}$
  - **Characteristic function:** $\phi(t) = \frac{\lambda}{\lambda - it}$

In general, the Exponential distribution describes the waiting time between two events (e.g., decay time, arrival time of next customer in a queue, time to next call at call center).

Property of exponential distribution: Lack of Memory

In general, lack of memory indicates that "the waiting time for the occurrence of an event does not depend on how long has passed up to now".

- ⇒ past does not influence the future.
- **More formally:**

$$P(T > t + \Delta t | T > t) = P(T > \Delta t)$$

- In practice, this means that "the probability that the event takes longer than $t + \Delta t$ given that we already waited t is the same as the probability that it takes longer than Δt starting from 0".
  - ⇒ i.e., the fact that we already waited t, does not change the probability of waiting another Δt time.
- **The exponential distribution has this property ⇒ Exponential is memoryless.**

The exponential distribution has this property → **Exponential is memoryless**

Imagine T ~ Exponential(λ), then:

$$P(T > t + \Delta t | T > t) = \frac{P(T > t + \Delta t \cap T > t)}{P(T > t)} = \frac{P(T > t + \Delta t)}{P(T > t)} = \frac{e^{-\lambda(t+\Delta t)}}{e^{-\lambda t}} = e^{-\lambda \Delta t}$$

Note that:

$$P(T > t) = 1 - F(t) = 1 - 1 + e^{-\lambda t} = e^{-\lambda t}$$

- **Example:** Imagine our survival time is exponentially distributed. Then:

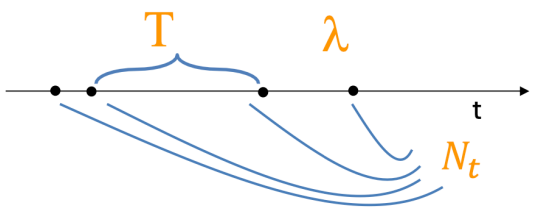
$$P(T > 90 + 5 | T > 90) = P(T > 5)$$

  - ⇒ Probability that a 90-year-old person survives 5 years is the same as a newborn!

Poisson and Exponential Relationship

Poisson and Exponential distributions model different aspects of the same process.

- On one side: counting event occurred in $\Delta t \rightarrow$ Poisson.
- On the other: waiting time between events $\rightarrow$ Exponential.
- So, are these R.V.s related?
- **Let** T be the waiting time between two successive events.
- **Let** N be the number of events occurring starting from $T = t$.
- What R.V. describes the waiting time T?



- We can work on the cdf:

$$F(t) = P(T \leq t)$$

- Now, let us focus on the event $A =$"no event up to $T = \Delta t$".
  - $P(A) = P(T > \Delta t) = 1 - F(\Delta t)$.
  - This also means that $N = 0$, so $P(A) = P(N = 0)$.
  - But N is a counting process, with intensity $\lambda = r\Delta t$, so $N \sim Poisson(\lambda)$.
  - Hence:

$$P(A) = P(N = 0) = \frac{(\lambda\Delta t)^0}{0!}e^{-\lambda\Delta t} = e^{-\lambda\Delta t} = 1 - F(\Delta t)$$

- $\Rightarrow F(\Delta t) = 1 - e^{-\lambda\Delta t} \Rightarrow$ Exponential cdf!

Interpretation:

- The Poisson distribution counts **how many events occur** in a fixed time.
- The Exponential distribution measures **how long you wait** between events.
- The rate λ connects the two:
  - In the Poisson distribution, λ is the average number of events per unit time.
  - In the Exponential distribution, λ is the rate of occurrence (or intensity) of events.

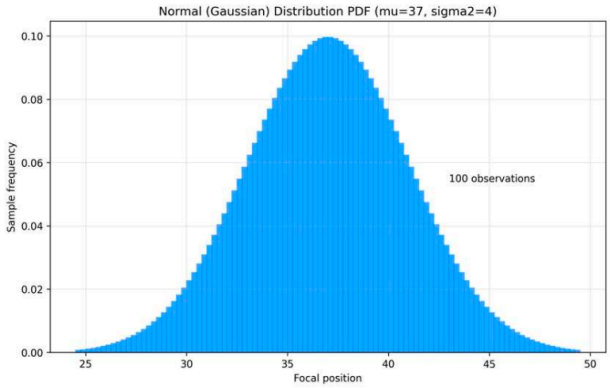
5. Normal Distribution

Example: Optical aberrations and lens defects.

- We focus a beam of light through a lens that, due to imperfections, produces random deviations from the ideal focal point μ
- Deviations are symmetrical, i.e., they are equally likely to be to the left or right of the ideal focal point.
- Small deviations are common, while large deviations are rare.
- **Question:** What is the actual focal point's position?
- **Mathematical Formulation:**
  - Let X be the random variable for the position of the actual focal point.
  - **Sample Space:** $S = (-\infty, \infty)$, meaning X is continuous.
- **Probability Density Function (pdf):**
  - The distribution is bell-shaped and centered around (\mu), with the following pdf:

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

where μ is the mean, and σ is the standard deviation.



- **Cumulative Distribution Function (cdf):**
  - Starts at 0, gradually increases, accelerates at a point, then slowly approaches 1 as $x \to \infty$.
  - No closed-form solution, requires numerical computation.
- **Expected Value:**

$$E(X) = \mu$$

- **Variance:**

$$V(X) = \sigma^2$$

- **Characteristic Function:**

$$\phi(t) = e^{i\mu t - \frac{1}{2}\sigma^2 t^2}$$

General Properties of the Normal Distribution

- **Standard Normal Distribution:** Any normal distribution $X \sim N(\mu, \sigma^2)$ can be transformed to the standard normal distribution $Z \sim N(0,1)$ by:

$$Z = \frac{X - \mu}{\sigma}$$

  - This distribution has mean 0 and variance 1, useful for standard computations.
- **Probability Intervals:**
  - 68% of data falls within $\pm 1\sigma$,
  - 95% within $\pm 2\sigma$,
  - 99.7% within $\pm 3\sigma$,
  - 99.99994266% within $\pm 5\sigma$, setting a high threshold for new discoveries.

Law of Large Numbers

The **Law of Large Numbers** states that the average of independent and identically distributed (i.i.d.) random variables converges to **their** expected value as the sample size increases.

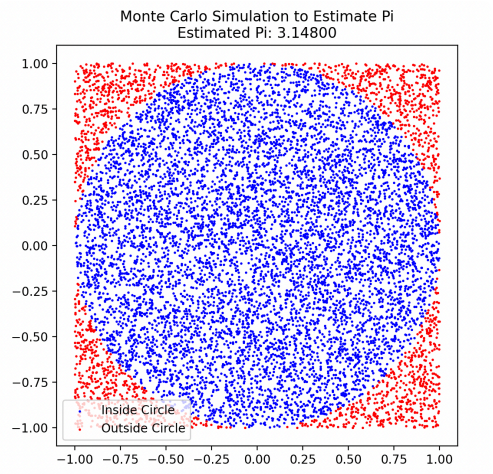
- Let ($X_1, X_2, \ldots, X_n$) be i.i.d. random variables:
  - $E(X_j) = \mu$ and $V(X_j) = \sigma^2$ for ($j = 1, \ldots, n$).
  - Define $\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$

    - $\boldsymbol{E(\bar{X}_n) = E\left(\sum_{j=1}^{n} \frac{X_j}{n}\right) = \frac{1}{n}E\left(\sum_{j=1}^{n} X_j\right) = \frac{1}{n}\sum_{j=1}^{n} E(X_j) = \frac{1}{n}n\mu = \mu}$
    - $\boldsymbol{V(\bar{X}_n) = V\left(\sum_{j=1}^{n} \frac{X_j}{n}\right) = \frac{1}{n^2}V\left(\sum_{j=1}^{n} X_j\right) = \frac{1}{n^2}\sum_{j=1}^{n} V(X_j) = \frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n}}$
  - Hence if we take the limit for n → ∞:
    - $\lim_{n\to\infty} E(\bar{X}_n) = \mu$ and $\lim_{n\to\infty} V(\bar{X}_n) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$

    → $\bar{X}_n$ converges to a constant value μ as n increases, independently of the initial distributions

The law of large numbers iv very powerful

My estimation of pi



the thing is that, the more points I generate, the more precise I'll get, this means that the error on the prediction will decrease:

- The x is generate randomly in uniform distribution
- The y is generate randomly in uniform distribution

To estimate the π i counted the number inside the circle. This corresponds to do the average of a quantity that is 0 if the point is inside the circle and 0 otherwise, that is randomly distributed.
This means that the limit on the average of this quantity is

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} I_i = \mathbb{E}[I] = \frac{\pi}{4}$$

Why the expected value of I is $\pi/4$?

- This is because

Thus, the probability that a randomly chosen point lies inside the circle is the ratio of the areas:

$$P(\text{point inside circle}) = \frac{A_{\text{circle}}}{A_{\text{square}}} = \frac{\pi}{4}$$

4. Expected Value of I:

Since $I = 1$ when the point is inside the circle and $I = 0$ when it is outside the circle, the expected value of I (i.e., the average value of I over many trials) is simply the probability that a point is inside the circle. Therefore:

$$\mathbb{E}[I] = P(\text{point inside circle}) = \frac{\pi}{4}$$

That's it.

Central Limit Theorem (CLT)

The **Central Limit Theorem (CLT)** asserts that the sum or average of a large number of i.i.d. random variables follows a normal distribution, regardless of the original distribution.

Intuition

- A process influenced by many independent factors can be seen as the sum of those factors.
- With many factors, the distribution of the sum or average approaches a normal distribution.

Formal Statement

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with $E(X_j) = \mu$ and $V(X_j) = \sigma^2$.
  - Define $\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$.
  - The standardized form $Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ approaches $N(0,1)$ as $n \to \infty$.

Notes

- For finite n, approximately valid to the extent that the **fluctuation of the sum is not dominated by one (or few) terms**
  Examples:
  - Good: velocity component vx of air molecules
  - Ok: total deflection due to multiple Coulomb scattering (rare large angle deflections give non-Gaussian tail)
  - Bad: energy loss of charged particle traversing thin gas layer (rare collisions make up large fraction of energy loss, cf. Landau pdf)
- **For finite *n*, if $X_j \sim N$**, then the pdf of S_n **is exactly Gaussian**. In fact:

$$\Phi_{S_n}(t) = \prod_j \Phi_{X_j}(t) = \prod_j e^{i\mu_j t - \frac{1}{2}\sigma_j^2 t^2} =$$

$$= e^{\Sigma(i\mu_j t - \frac{1}{2}\sigma_j^2 t^2)} = e^{i\Sigma(\mu_j)t - \frac{1}{2}\Sigma(\sigma_j^2)t^2}$$

μ_n $\quad$ σ_n^2

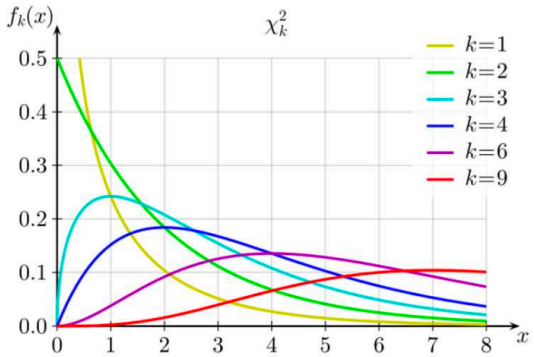
11

---

6. Chi-Squared Distribution

- **Definition:** Chi-squared distribution describes a positive definite random variable, obtained as the sum of squared independent standard Gaussian.
  - **Applications:** Used in hypothesis testing and confidence intervals for variance.
- **Properties:**
  - $X \sim \chi^2(\nu)$, where ν is the degrees of freedom, representing the number of squared components.
  - **Sample Space:** $S = [0, \infty)$
- **Probability Density Function (pdf):**

$$f(X = x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2}$$



pdf

- **Expected Value:** $E(X) = \nu$
- **Variance:** $V(X) = 2\nu$

---

t-Student Distribution

- **Definition:** Similar to the Gaussian distribution but with heavier tails, suitable for smaller sample sizes.
  - **Applications:** Used in hypothesis testing and confidence intervals for mean.
- **Properties:**
  - $X \sim t(\nu)$, where ν is the degrees of freedom.
  - As $\nu \to \infty$, t approximates $N(0, 1)$
  - **Sample Space:** $S = (-\infty, \infty)$
- **Probability Density Function (pdf):**

$$f(X = x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- **Expected Value:** $E(X) = 0$ for $\nu > 1$; otherwise undefined.
- **Variance:** $V(X) = \frac{\nu}{\nu-2}$ for $\nu > 2$; infinite for $1 < \nu \leq 2$

How do multiple factors and their combinations influence a process?

Outline of Topics

1. **Modeling processes with multiple influencing random factors**: How multiple random variables can be used to describe complex systems.
2. **Multivariate Random Variables**: Joint, marginal, and conditional probability distributions.
3. **Transformations of Random Variables**: Techniques for deriving distributions of functions of random variables.

Introduction to Multivariate Random Variables

When dealing with systems influenced by more than one random factor, we use multivariate random variables. Instead of analyzing each variable in isolation, multivariate analysis examines their combined behavior.

Example: Weather Forecasting Model

Consider a model predicting weather in a coastal city. In this case, two observable random variables could define the daily conditions:

- (X): Daily maximum temperature in degrees Celsius (°C).
- (Y): Relative humidity percentage (%).

Our question: **What is the probability of observing a particular combination of temperature and humidity?**

Since both (X) and (Y) influence weather conditions, we cannot consider them independently. Instead, we use a **Joint Probability Distribution** to describe the likelihood of specific combinations of temperature and humidity.

JOINT PROBABILITY DISTRIBUTION

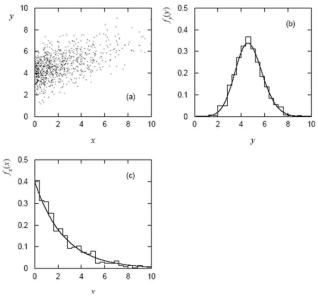
For two random variables (X) and (Y), the **joint probability distribution** (f(X, Y)) quantifies the probability of simultaneously observing particular values of both (X) and (Y). For instance, if (f(25, 80)) represents the probability of a 25°C day with 80% humidity, then we can formally express this as:

$$P(A \cap B) = \iint f(x, y)\, dx\, dy$$

MARGINAL DISTRIBUTION

The **marginal distribution** focuses on the probability of observing values of one variable while ignoring the influence of the other. For example, to find the probability of a specific temperature regardless of humidity, we integrate over all possible values of (Y):

$$P(A) = \int f(x, y)\, dy = f(x)$$



Marginal pdf: intuition

Intuition:

Think of marginalization as a projection of the joint pdf onto one of the axis

Conditional Distribution

If we want to know the probability of humidity (Y) given a fixed temperature (X), we use the **conditional distribution**:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \Rightarrow f_{Y|X}(y|x) = \frac{f(x, y)}{f(x)}$$

This ratio normalizes the joint probability to ensure that probabilities sum up to one over the conditional distribution's domain.

Notes:

Conditionals can be expressed as:

- $h(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$
- $g(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}$

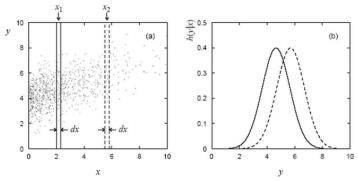
Hence Bayes theorem becomes:

$$g(x|y) = \frac{h(y|x) f_X(x)}{f_Y(y)}$$

Also, if $X \perp Y \rightarrow f(x, y) = f_X(x) f_Y(y)$

y?

Conditional pdf: intuition



Intuition:

Fix a conditioning R.V., e.g. X

Study the joint distribution over Y only in at the fixed value of X

To get probabilities, divide by the marginal of X to ensure normalization, e.g., $\int_Y h(y|x)dy = 1$

Multivariate Moments

Multivariate moments extend the concept of moments (like mean and variance) to cases involving two or more random variables. These moments help quantify the joint behavior of variables.

Mixed Moments

For two random variables X and Y with means μ_X and μ_Y, respectively, **mixed moments** of order (m, n) are defined as:

$$V_{m,n} = E[(X - \mu_X)^m (Y - \mu_Y)^n]$$

COVARIANCE

Most used is: **Covariance**

The **covariance** between X and Y, a commonly used mixed moment, measures how much X and Y vary together:

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - E[X]E[Y]$$

If $\text{Cov}(X, Y) > 0$, X and Y tend to increase together; if $\text{Cov}(X, Y) < 0$, one increases as the other decreases.

Correlation Coefficient

The **correlation coefficient** ρ standardizes covariance to a value between -1 and 1, representing the strength and direction of a linear relationship:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- $\rho = 1$: Perfect positive correlation
- $\rho = -1$: Perfect negative correlation
- $\rho = 0$: No linear correlation

Note: if $X \perp Y \rightarrow f(x, y) = f_X(x) f_Y(y)$. This implies that:

$$E[XY] = \int\int xy\, f(x, y)\, dx\, dy = \int_X x f(x)\, dx \int_Y y f(y)\, dy = E(X)E(Y)$$

→ Substituting in Covariance formula: $\text{COV}[X, Y] = 0$
→ This in turn implies $\rho_{XY} = 0 \rightarrow$ uncorrelated

> Note: the inverse is not always true, i.e. uncorrelated **DOESNT MEAN** independent.
> Clear.

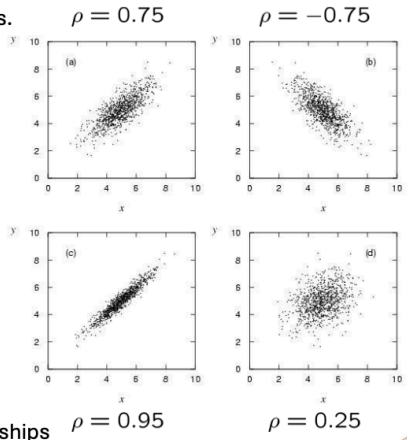
Covariance Matrix

In the multivariate context, we use a **covariance matrix** to summarize the variances and covariances of a set of random variables. For a vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$:

- Diagonal elements represent variances, i.e., $V_{ii} = \text{Var}(X_i)$
- Off-diagonal elements represent covariances, i.e., $V_{ij} = \text{Cov}(X_i, X_j)$

Correlation: intuition

Correlation quantifies the **linear** relationship between two variables.



- The correlation coefficient ρ ranges from -1 to +1:
  - ρ = +1 indicates perfect positive correlation (a and c)
  - ρ = -1 indicates perfect negative correlation (b)
  - ρ = 0 indicates no linear correlation (d)
- As |ρ| increases, the scatter plot becomes more tightly clustered around a line:
  - Strong positive/negative correlation (cases *a, b*): as x increases, y tends to increase/decrease (respectively)
  - Very strong positive correlation (ρ = 0.95): Tight linear relation
  - Weak positive correlation (ρ = 0.25): Loose, scattered relation

Key points:

- Correlation measures **strength and direction of linear** relationships
- Non-linear relationships may have low ρ despite strong dependencies (*e.g. y=x^2*)
- Correlation **does not imply causation** → check spurious correlations website for funny ones!

3

- Example

Dependent variable can actually have null correlation coefficient!

- Consider two random variables X, Y such that:
  - $X \sim Unif(-1, 1)$
  - $Y = X^2$
- Clearly, they are dependent as Y is a function of X → knowing X gives us direct access to the exact value of Y
- However, it can be shown they are uncorrelated

Exercise

Show that X, Y are uncorrelated

Note that $E[X] = \frac{a+b}{2} = \frac{-1+1}{2} = 0$

→ $COV(X,Y) = E[XY] - E[X]E[Y] = E[XY] = E[XX^2] = \int_{-1}^{1} \frac{x^3}{2} dx = 0$

Multivariate Gaussian Distribution

One important example of multivariate distributions is the **multivariate Gaussian (normal) distribution**. Suppose $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ consists of i.i.d. normal variables, each with mean μ and variance σ^2. Then $\mathbf{X}$ has a **multivariate Gaussian distribution**:

$$\mathbf{X} \sim \mathcal{N}(\mu, V^2)$$

where:

- μ is the vector of means for each variable.
- V is the covariance matrix.

The probability density function (pdf) for a multivariate normal is:

$$f(\mathbf{X} = \mathbf{x}; \mu, V) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T V^{-1}(\mathbf{x}-\mu)}$$

where $|V|$ is the determinant of V

Special case, n=2 components:

$$f(x_1, x_2, ; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$
$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right\}$$

where $\rho = \text{cov}[x_1, x_2]/(\sigma_1\sigma_2)$ is the correlation coefficient

---

Functions of Random Variables

Transforming random variables often yields new variables with their own distributions. Suppose $Y = g(X)$, where g is a function of X. The pdf of Y is given by:

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| \quad \text{where } x = g^{-1}(y)$$

- Example

Let X be a uniform R.V. such that $X \sim U(0,1)$. Then consider its transformation $Y = g(X) = 2x$.
What is the pdf of Y?

Solution

Since g is monotonic, then: $f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$

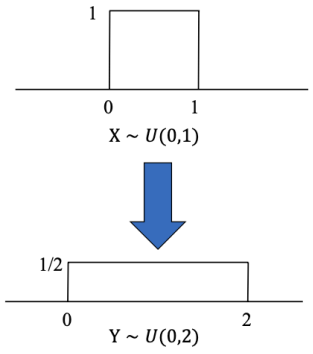
Hence to compute the pdf of Y we need:

- $g^{-1}(y) = \frac{y}{2}$
- $\frac{dg^{-1}(y)}{dy} = 1/2$

By substituting: $f_Y(y) = \frac{1}{2}U_{(0,1)}(g^{-1}(y)) = \frac{1}{2}U_{(0,1)}\left(\frac{y}{2}\right) = \frac{1}{2}$

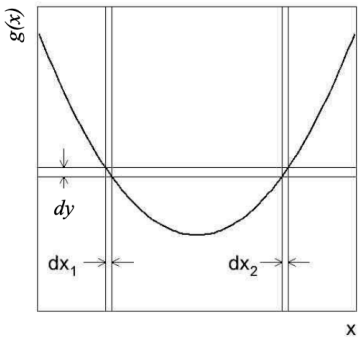
Note: the range also changes when applying g:

- g(0)=0, g(1)=2 → $Y \sim U(0,2)$



What if the transformation is not monotonic? i.e. **what if not unique inverse?**

- In general, $f_Y(y)dy = \int_{dS} f_X(x)dx$
- If g has not a unique inverse, then we simply include in dS all dx intervals corresponding to dy, *i.e.* $dS = [dx_2] \cup [dx_1]$



Example

Let $Y = g(X) = x^2$. It follows that:

- $g^{-1}(y) = \pm\sqrt{y}$
- $\frac{dg^{-1}(y)}{dy} = \pm\frac{1}{2}y^{-\frac{1}{2}} \rightarrow dx = \frac{dy}{2\sqrt{y}}$

Hence: $dS = \left[\sqrt{y}, \sqrt{y} + \frac{dy}{2\sqrt{y}}\right] \cup \left[-\sqrt{y} - \frac{dy}{2\sqrt{y}}, -\sqrt{y}\right]$

And $f_Y(y) = \frac{f_X(\sqrt{y})}{2\sqrt{y}} + \frac{f_X(-\sqrt{y})}{2\sqrt{y}}$

In general, let X be a R.V. with probability function $f(x)$ and let $Y = g(X)$ be its transformation. Assuming $g(x)$ is invertible (but the inverse is not necessarily unique), the probability function of Y can be derived in 3 steps:

- For each y, find the set $A_y = \{x: g(x) \leq y\}$
- Find the cdf, $F_Y(y)$:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(\{x; g(x) \leq y\}) = \int_{A_y} f_X(x)dx$$

- Finally, derive the pdf by differentiating:

$$f_Y(y) = F'_Y(y)$$

- Example

  Let X be a R.V. with $f_X(x) = e^{-x}$ for $x > 0$. Let $Y = g(X) = \ln(X)$ be its transformation. What is the pdf of Y?

  Solution

  - Observe that $g^{-1}(y) = e^y$
  - Therefore: $A_y = \{x: \ln(x) < y\}$, or equivalently: $A_y = \{x: x < g^{-1}(X)\}$
  - Find the cdf:

$$F_Y(y) = P(Y \leq y) = P(\ln(X) \leq y) = P(X \leq e^y) = F_X(e^y) = 1 - e^{-e^y}$$

  - Finally, the pdf of Y is:

$$f_Y(y) = \frac{dF_Y(y)}{dy} = 0 - \left(-e^y e^{-e^y}\right)1 = e^y e^{-e^y}$$

  Alternatively

  - Note that g() is monotone, so:

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right| = e^{-e^y}1e^y$$

Functions of Many Random Variables

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random vector with a known joint probability density function $f_{\mathbf{X}}(\mathbf{x})$. Let $\mathbf{Y} = g(\mathbf{X})$ be a new random vector where each component of $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$ is a function of $\mathbf{X}$.

- Then the pdf of $\mathbf{Y}$ can be expressed as:

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})\left|\det\frac{d\mathbf{x}}{d\mathbf{y}}\right|$$

  where
  - $\mathbf{x} = g^{-1}(\mathbf{y})$ is the inverse transformation (assuming it exists)
  - $\frac{d\mathbf{x}}{d\mathbf{y}}$ is the Jacobian matrix of partial derivatives
  - we take the determinant of the Jacobian.

Examples

Let X, Y be two random variables with a known joint probability density function $f_{X,Y}(x, y)$. Let $Z = X + Y$ be a new random variable. What is the distribution of Z?

- In general:

$$f_Z(z) = \int f_{X,Y}(x, z - x)\, dx$$

- However, if X and Y are independent, we can further decompose:

$$f_Z(z) = \int f_X(x) f_Y(z - x)\, dx$$

  that is the convolution formula.

Error Propagation

Let X be a random variable representing the measurement of a physical quantity, and let $Y = g(X)$ be its transformation. Suppose $V(X)$ is known, which quantifies the error on the measurement of X. What is the variance (error) of Y?

In principle, we could compute the variance of $g(X)$ analytically by exploiting the definition. However, this is often impractical:

- We may not know the distribution of X
- Calculation could be too complex for direct computation.

Alternatively, we can use the error propagation formula for approximating $V(Y)$:

- Use first-order Taylor series expansion of $g(X)$ around $E(X) = \mu$

$$g(X) \approx g(\mu) + g'(\mu)\,(X - \mu)$$

- Then:

$$V(Y) = E\left[g'(\mu)^2\,(X - \mu)^2\right] = g'(\mu)^2 V(X)$$

  - This means the variance of Y is proportional to the variance of X, scaled by the square of the derivative of $g(X)$ computed at μ.

In general, if $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and $\mathbf{Y} = g(\mathbf{X})$, then:

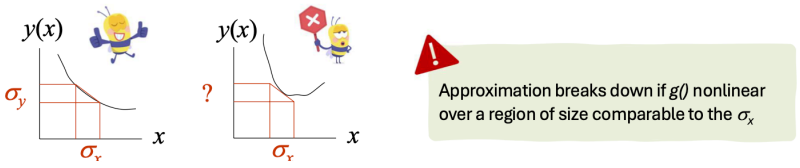
$$V(Y) = \left(\frac{dg}{dX_1}\right)^2 V(X_1) + \cdots + \left(\frac{dg}{dX_n}\right)^2 V(X_n) + 2\sum_{i \neq j} \frac{dg}{dX_i}\frac{dg}{dX_j}\,\mathrm{Cov}(X_i, X_j)$$

- Note: no assumptions about the distributions of X_i.

Limitations

While convenient in practice, several assumptions must hold for error propagation to be effective:

- $g(X)$ is smooth and can be well-approximated by a linear expansion.
- Uncertainties in the random variable are relatively small, so higher-order terms in Taylor expansion can be neglected.
- Variances and covariances of X are known (or can be estimated).



Examples

Example 1

Let X_1, X_2 be two random variables and define $Y = X_1 + X_2$ and $Z = X_1 X_2$.

- By applying error propagation formulas, we get:

$$V(Y) = V(X_1) + V(X_2) + 2\,\mathrm{Cov}(X_1, X_2)$$

$$V(Z) = \left(\frac{\partial Z}{\partial X_1}\right)^2 V(X_1) + \left(\frac{\partial Z}{\partial X_2}\right)^2 V(X_2) + 2\frac{\partial Z}{\partial X_1}\frac{\partial Z}{\partial X_2}\,\mathrm{Cov}(X_1, X_2)$$

- If X_1 and X_2 are uncorrelated:
  - Add errors quadratically for sum, Y (or difference).
  - Add relative errors quadratically for product, Z (or ratio).
  - Do not apply when correlations are present.

Exercise: Derive the formula for $V(Z)$.

If $\mu_{X_1} = E(X_1)$ and $\mu_{X_2} = E(X_2)$, the formula becomes:

$$V(Z) = \mu_{X_2}^2 V(X_1) + \mu_{X_1}^2 V(X_2) + \mu_{X_1}\mu_{X_2}\,\mathrm{Cov}(X_1, X_2)$$

Example 2

Let X_1, X_2 be independent continuous $Uniform(0,1)$ random variables. Find the density of $Y = g(X_1, X_2) = X_1 + X_2 \cdot X_2$.

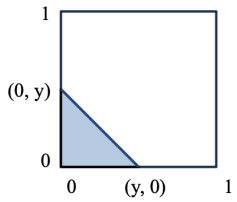
Solution:

- Define the cumulative distribution function $F_Y(y)$:

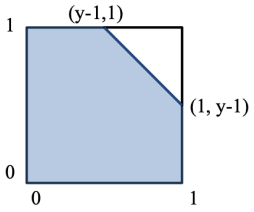
$$F_Y(y) = P(Y \le y) = P(g(X_1, X_2) \le y) = P((x_1, x_2) : g(x_1, x_2) \le y) = \int\int f_{X_1, X_2}(x_1, x_2)\,dx_1\,dx_2$$

Solution

- $F_Y(y) = P(Y \le y) = P(g(X_1, X_2) \le y) = P(\{(x_1, x_2)\}: g(x_1, x_2) \le y) = \int \int_{A_y} f(x_1, x_2) dx_1 dx_2$
- What is A_y?
  - $0 \le y < 1$
  - $1 \le y \le 2$



$0 \le y < 1$



$1 \le y \le 2$

Statistical inference: estimators and information

Outline

- How can we learn from data?
- Estimators
- Maximum Likelihood

How can we learn from data?

Statistical inference, or «learning», is the process of extracting knowledge about a phenomenon from its own data.

- What is the distribution of the data? And what are its properties?
- Typically, we cannot observe the whole population (limited time, resources, or process not fully observable).
- We resort to a «sample» instead and only observe limited evidence/data.

Derived questions:

- How to derive properties or models based on partial information?
- How to quantify uncertainty?
- How to test hypotheses and make predictions?

Parametric vs non-parametric inference

There are several approaches to statistical inference, depending on assumptions. A key distinction is between parametric and non-parametric inference.

Parametric
We assume the process is described by a function with a finite set of parameters:

$$X \sim \mathcal{F}(x;\theta), \mathcal{F} = f(x;\theta) : \theta \in \Theta$$

Where:

- $\mathcal{F}$ is a family of functions parametrized by θ.
- θ is the fixed (there is a true value that describes it, is not random) but unknown (vector of) parameters.
- Θ is the parameter space, i.e., all allowed values for θ.

Goal: Inference about distribution parameters θ.

- More powerful if assumptions hold :)
- Incorrect if assumptions are violated :(

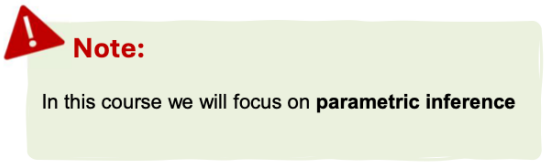
Non-parametric
We do not assume a specific functional form for $\mathcal{F}$. It is modelled by a non-finite number of parameters.
This means that $\mathcal{F}$ is modelled by a non-finite number of parameters

- we do not make assumptions on $\mathcal{F}$
- $\mathcal{F}$ has no parameter

It's more flexible and we allow a more number of parameters.

Goal: inference about data characteristics, e.g. median as central tendency

- More flexible, robust to outliers/deviations :)
- More complex and less powerful than valid parametric alternatives :(



Parametric vs non-parametric inference: example

Measuring the speed of sound in air:
We take 50 measurements under identical conditions.

- **Parametric**
  Assume measurements follow a Gaussian distribution:

$$X \sim \mathcal{F}(x;\theta) = f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} : \mu \in \mathbb{R}, \sigma > 0$$

  Estimate the parameters μ, σ. Inference is based on $N(\mu,\sigma)$.
- **Non-parametric**
  No assumption on underlying distribution $\mathcal{F}$. Use median and interquartile range (IQR) for central tendency and spread.
  - Use median as measure of central tendency
  - Use interquartile (IQR) range for spread -> our description depends on summary statistics of observed data: median and IQR

Parameters of interest vs nuisance parameters

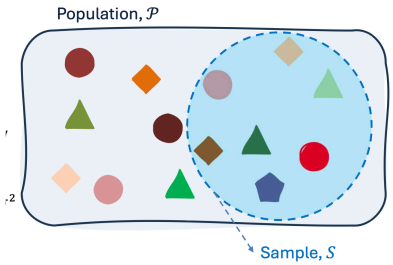
Imagine we know that we can describe the distribution F with a set of parameters, now more than 1: $X \sim \mathcal{F}(x;\theta = \{\alpha,\beta\})$

In all the parameters we chose parameters in which we are interested in and others that we think are not that interesting.

- α: parameters of interest (e.g., μ).
- β: **nuisance parameters** (e.g., σ), which determine the shape of $f(x;\mu,\sigma)$ but are not of interest.

Example: Gaussian distribution but we are interested in the estimation of the expected value of the population, not how spread it is.

From population to sample



Typically we cannot observe the whole population: Hence we resort to sampling

- ***Population***: entire group of individuals/entities under study
  - Often too large/impractical to observe
  - E.g. all electrons in the universe, all possible decays
  - Population quantities are referred to with Greek letters: θ, μ, σ
- ***Sample***: subset of the population
  - Useful to make inference without observing all population
  - E.g. electrons revealed in a specific experiment
  - Sample quantities are referred to with Latin letters: $\bar{x}$, s^2
  - Must be representative of the population -> different sampling methods ensure different properties (not covered here)
- ***Statistical units***: individual elements of the population/sample
  - Basic entities on which measurements/observations are made
  - E.g. single electron

Inference -> retrieving information about the population starting from a sample

Statistical inference: sub-problems

1. Estimating parameters (point estimates, interval estimation).
   - I know the model and I want to estimate its parameters
2. Hypothesis testing.
   - Compare two models/hypotheses
3. Goodness-of-fit.
   - Measure how well a model/hypothesis fit the data (a model VS all the others)

Data statistics and estimators

How do we estimate our parameters of interest, θ?

We define a statistic as a generic function of data: $t = t(\vec{X}) = t(X_1, \ldots, X_n)$.

- Examples:
  Parameters of interest: expected value (μ), variance (σ^2).
  Related statistics: sample mean $(\bar{X})$, sample variance (s^2).
- **How do we choose good statistics as estimators?**

Bias and precision

We are conducting a statistic and we are conducting some inference assuming that x is distributed in a F distribution.

Bias: The bias of a statistic $t(X)$ as an estimator for θ is:

$$B(t(X)) = E(t(X)) - \theta$$

- The bias measures how close, on average, the statistic is to the true parameter value: Measures the accuracy of the statistic.
- Ideally, we would like the bias to be as low as possible
- A statistic with 0 bias is called **unbiased** estimator of θ

Precision: The precision of a statistic $t(X)$ is:

$$\text{precision}(t(X)) = \frac{1}{V_\theta[t(X)]}$$

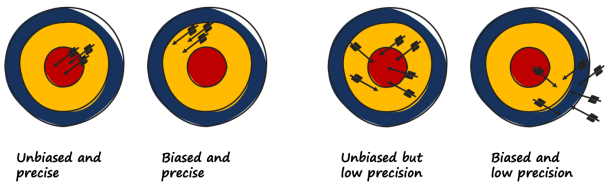
- The precision measures the **dispersion** of the statistic around its expected value.
- Measures the **variability** of the statistic.
- A statistic with highest precision is called efficient estimator of θ:

  $$V_\theta[t(X)] \leq V_\theta[t^*(X)]$$

  for any t^*.

Accuracy vs precision

- **Accuracy:** How close the estimator is to the true value.
- **Precision:** How variable the estimator is.



Mean Squared Error (MSE)

A measure of the quality of an estimator is given by the MSE:

$$MSE(t(X)) = E_\theta\left[(t(X)-\theta)^2\right] = V(t(X)) + B(t(X))^2$$

- Trade-off between bias and variance.
  - Derivation:

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathbb{E}_{\hat{\theta}}\left[(\hat{\theta}-\theta)^2\right] \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\mathbb{E}_{\hat{\theta}}[\hat{\theta}]+\mathbb{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)^2\right] \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\mathbb{E}_{\hat{\theta}}[\hat{\theta}]\right)^2 + 2\left(\hat{\theta}-\mathbb{E}_{\hat{\theta}}[\hat{\theta}]\right)\left(\mathbb{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right) + \left(\mathbb{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)^2\right] \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\mathbb{E}_{\hat{\theta}}[\hat{\theta}]\right)^2\right] + \mathbb{E}_{\hat{\theta}}\left[2\left(\hat{\theta}-\mathbb{E}_{\hat{\theta}}[\hat{\theta}]\right)\left(\mathbb{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)\right] + \mathbb{E}_{\hat{\theta}}\left[\left(\mathbb{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)^2\right] \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\mathbb{E}_{\hat{\theta}}[\hat{\theta}]\right)^2\right] + 2\left(\mathbb{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)\mathbb{E}_{\hat{\theta}}\left[\hat{\theta}-\mathbb{E}_{\hat{\theta}}[\hat{\theta}]\right] + \left(\mathbb{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)^2 && \mathbb{E}_{\hat{\theta}}[\hat{\theta}]-\theta = \text{const.} \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\mathbb{E}_{\hat{\theta}}[\hat{\theta}]\right)^2\right] + 2\left(\mathbb{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)\left(\mathbb{E}_{\hat{\theta}}[\hat{\theta}]-\mathbb{E}_{\hat{\theta}}[\hat{\theta}]\right) + \left(\mathbb{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)^2 && \mathbb{E}_{\hat{\theta}}[\hat{\theta}] = \text{const.} \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\mathbb{E}_{\hat{\theta}}[\hat{\theta}]\right)^2\right] + \left(\mathbb{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)^2 \\
&= \mathrm{Var}_{\hat{\theta}}(\hat{\theta}) + \mathrm{Bias}_{\hat{\theta}}(\hat{\theta},\theta)^2
\end{aligned}$$

> **Bias/Variance trade-off:**
> In practice, we cannot minimize both bias and variance simultaneously → **compromise between accuracy and precision!**

Consistency and sufficiency

Let $X \sim \mathcal{F}(x;\theta), \mathcal{F}(f(x;\theta):\theta\in\Theta)$ and $t(X)$ be a statistic. The statistics $t(X)$ could have the property of

Consistency: A statistic $t(X)$ is consistent if:

$$t(X) \to \theta, \quad \text{as } n \to \infty$$

Sufficiency: A statistic $t(X)$ is sufficient for θ if:

$$P(X|t(X),\theta) = P(X|t(X))$$

- all information about θ is already contained in $t(X)$
  - it embeds all useful information for the parameters of interest

Summary of data statistics properties

Let $X \sim \mathcal{F}(x;\theta), \mathcal{F}(f(x;\theta):\theta\in\Theta)$ and $t(X)$ be a statistic. Then $t(X)$ will be good as an estimator for θ when it satisfies the following properties:

- **Unbiasedness:** On average, does the statistic hit the true parameter value? $E(t(X)) = \theta$
  - In practice, we want low bias (ideally 0), which means low systematic error: $b = E(t(X)) - \theta$
- **Efficiency:** How much does it vary around the true value? $V(t(X)) \leq V(t^*(X))$, for any t^*
  - In practice, we want as little variability around θ as possible.
- **Consistency:** Does the statistic converge to the true value as the sample size increases?

$t(X) \to \theta$, that is:
$\lim_{n\to\infty} P(|t(X) - \theta| < \epsilon) = 1$
    - For $n \to \infty$, we have that the estimator converges to a fixed value equal to the true parameter (i.e., zero variance).
- **Sufficiency**: Does it embed all useful information for the parameters of interest?
$P(X|t(X), \theta) = P(X|t(X)) \forall \theta$
    - This means that all information about θ is already contained in $t(X)$.

Point Estimation

Imagine we know that $X \sim \mathcal{F}(x;\theta)$, $\mathcal{F}\{f(x;\theta) : \theta \in \Theta\}$. Point estimation revolves around providing a single "best guess" for a quantity of interest.
What can you 'best guess' are, for example:

- Parameter of a distribution $\mathcal{F}$
- The whole distribution, e.g., cdf/pdf of $\mathcal{F}$
- A regression function $r(X)$ assuming that $X \sim \mathcal{F}$
- A prediction of a future value X

Notes:

- By convention, we denote our estimate as $\hat{\theta}$
- θ is unknown but fixed, and is the true value.
- However, $\hat{\theta}$ is a random variable as it depends on data

How can we provide a point estimate for a parameter θ?

1. Method of moments
2. Maximum likelihood

1. Method of Moments

Let $X \sim \mathcal{F}(x;\theta)$, $\mathcal{F}\{f(x;\theta) : \theta \in \Theta\}$, where $\theta = \{\theta_1, \ldots, \theta_K\}$ is a K-dimensional vector of parameters. Then, the method of moments estimator $\hat{\theta}_n$ can be derived as follows:

- Define the j-th moment α_j, $1 \leq j \leq K$ as $\alpha_j = \alpha_j(\theta) = E_\theta\left[X^j\right] = \int x^j dF_\theta(x)$
- The corresponding sample moment $\hat{\alpha}_j$ will be: $\hat{\alpha}_j = \frac{1}{n}\sum_i X_i^j$
- Then $\hat{\theta}_n$ is defined as:

$$\alpha_1(\hat{\theta}_n) = \hat{\alpha}_1$$

$$\alpha_2(\hat{\theta}_n) = \hat{\alpha}_2$$

$$\ldots$$

$$\alpha_K(\hat{\theta}_n) = \hat{\alpha}_K$$

- System of K equations with K unknowns
  - In practice, we estimate each moment by its sample version

Example:
Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$.
Then: $\alpha_1 = E(X_1) = p$ and $\hat{\alpha}_1 = \frac{1}{n}\sum_i X_i$.
Therefore: $\hat{p}_n = \frac{1}{n}\sum_i X_i$

---

2. Likelihood function and Maximum Likelihood

Let $X \sim \mathcal{F}(x;\theta)$, $\mathcal{F}\{f(x;\theta) : \theta \in \Theta\}$, where $\theta = \{\theta_1, \ldots, \theta_K\}$ is a K-dimensional vector of parameters.
Given a random vector $\mathbf{X} = \{X_1, \ldots, X_n\}$ of i.i.d. random variables, we can write the joint density as:

$$P(\mathbf{X}) = f(\mathbf{X};\theta) = \prod_{i=1}^{n} f(X_i;\theta)$$

Then, the likelihood function is defined as:

$$L_n(\theta;\mathbf{X}) = \prod_{i=1}^{n} f(X_i;\theta)$$

- Nothing but the joint density as a function of the parameter θ instead of the data $\mathbf{X}$
  - $L_n : \Theta \to [0, \infty)$
- The likelihood is *not* a density function $\to$ does not integrate to 1

Describes how likely it is to observe given data $\mathbf{X}$ under $\mathcal{F}$ depending on a specific parametrization θ

- degree of agreement between observed data and $\mathcal{F}$ parametrization by θ

The method of **maximum likelihood** provides estimates that maximize the likelihood of the observed data:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L_n(\theta;\mathbf{X})$$

The basic assumption is that what we are observing is not rare

- *This assumption means that the observed data X is typical or representative of the underlying distribution.*
- Parameter estimates are retrieved through optimization of the likelihood function with respect to θ:
  1. Derive with respect to θ
  2. Set equal to zero and check which zeros are maximum points

Rewrite the Likelihood

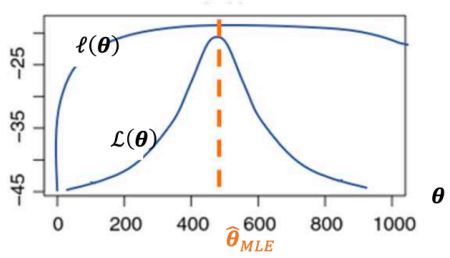
By definition, for independent random variables the likelihood can be written as a product:

$$L_n(\theta; X) = \prod_{i=1}^{n} f(X_i; \theta)$$

Taking the logarithm is convenient for differentiation
→ we typically work on log-likelihood instead:

$$\ell(\theta) = \log L_n(\theta; X) = \sum_{i=1}^{n} \log f(X_i; \theta)$$

- Note that optimal point do not change!



FORMAL: Maximum Likelihood estimation

More formally, let $X \sim F(x; \theta)$, $F(f(x; \theta) : \theta \in \Theta)$, where $\theta = (\theta_1, \ldots, \theta_k)$ is a K-dimensional vector of parameters. The maximum likelihood estimator of θ is defined as:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L_n(\theta; X)$$

In practice, to calculate $\hat{\theta}_{MLE}$ we can:

- Compute the log-likelihood: $\ell(\theta) = \log L_n(\theta; X)$
- Derive the log-likelihood: $S(\theta) = \frac{d\ell(\theta)}{d\theta}$
  → $S(\theta)$ typically referred to as the *score function*
- Set $S(\theta) = 0$ and solve for θ to find critical points
- Check which critical points correspond to a maximum → second derivative negative at $\hat{\theta}_{MLE}$

Example:
Let $X_1, \ldots, X_n \sim_{IID}$ Bernoulli(p). Then: $\mathcal{L}_n(p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$, where $S = \sum_i X_i$

- $\ell_n(p) = S \log p + (n - S) \log(1 - p)$
- $S(p) = \frac{d\ell(p)}{dp} = \frac{S}{p} - \frac{n-S}{1-p} = S - Sp - np + Sp = S - np$
- Solve $S(p) = 0 \rightarrow p = \frac{S}{n} = \frac{\sum_i X_i}{n}$ (Note: $\frac{d^2\ell(p)}{dp^2} = -n$ is always negative)

.

What is information in statistics?

The concept of information in statistics is related to the "knowledge" one can derive from data.

Example: You have performed an experiment and collected some data:

- What is the information data provide about the model?

In general, any measure of statistical information should satisfy some properties:

- The more data you collect, the more information you have. This improves parameter estimates or model understanding.
- Related to the parameters of interest.
  - Information should focus on what you're studying. Data irrelevant to your parameters should not increase information.
  - Should be related to precision ⇒ the larger the information, the better the precision.

Note: Data reduction typically implies information loss.
⇒ How to go from raw data to high-level summaries (reconstruction) minimizing information loss?

Shannon Information

Shannon's definition relates information to uncertainty.

Let X be a random variable with K possible outcomes $x_1,\ldots,x_K$, each with probability p_i. Then the information coming from observing an outcome x_i is defined as:

$$I(x_i) = \log\left(\frac{1}{p_i}\right) = -\log p_i, \text{ (base } b \text{ arbitrary)}$$

⇒ The smaller p_i, the higher the information.

> Example: talking on the phone
>
> 1. 1y old saying "Da" with probability 1
> 2. 3y old saying 500 words with probability $p_1, p_2, \ldots p_500$
>
> More info in the second

Starting from the above definition, Shannon information associated to the random process X is defined as its expected information:

$$H(X) = E[I] = -\sum_{i=1}^{K} p_i \log p_i \text{ (also called entropy)}$$

Intuition: The greater the entropy, the higher the information we gain by observing the random process.

Fisher Information

Fisher's definition links information to the knowledge a sample provides about an unknown parameter.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector representing a sample of n observations, and let $\mathcal{L}(\theta; \mathbf{X})$ be the corresponding likelihood function depending on the parameter θ. Then the information carried by the sample about θ is defined as:

$$I(\theta) = E\left[\left(\frac{\partial \ell(\theta; \mathbf{X})}{\partial \theta}\right)^2\right] = E\left[\left(\frac{\partial}{\partial \theta} \log \mathcal{L}(\theta; \mathbf{X})\right)^2\right]$$

Under quite general regularity conditions, we have: $E\left[\frac{\partial \ell(\theta; \mathbf{X})}{\partial \theta}\right] = 0$.
Hence, $I(\theta) = \mathrm{Var}\left(\frac{\partial \ell(\theta; \mathbf{X})}{\partial \theta}\right)$.

⇒ Fisher information is related to the variance of the score function.

Properties:

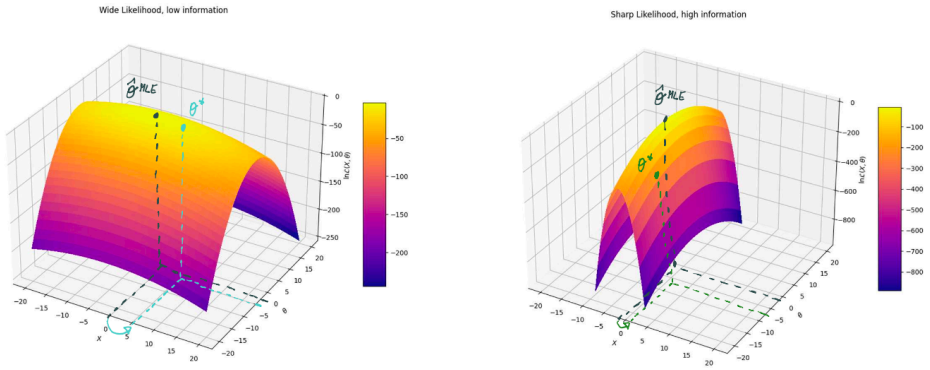
- Fisher information is non-negative.
- It plays a central role in the Cramér-Rao lower bound, which provides a lower bound for the variance of unbiased estimators.

In addition, if $\ell(\theta; \mathbf{X})$ is also twice differentiable with respect to θ, then:

$$I(\theta) = -E\left[\frac{\partial^2 \ell(\theta; \mathbf{X})}{\partial \theta^2}\right]$$

⇒ Fisher information is also linked to the curvature of the likelihood.

Intuition: Higher Fisher information implies that the data provides more information about the parameter, resulting in a smaller variance of the estimator.



Cramér-Rao Theorem (also Rao-Cramér-Frechet or RCF bound)

This theorem is a powerful tool that sets a lower bound for the variance of unbiased estimators for a parameter θ.

- Let $\hat{\theta}$ be an unbiased estimator for a parameter θ.
- Let $f(X, \theta)$ be the probability distribution of the data X.

- Then the Cramér-Rao theorem shows that:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

⇒ The inverse of Fisher information is the lower bound for the variance of any unbiased estimator of θ.

Note: This provides a reference setting for evaluating the efficiency of an estimator:

$$\text{efficiency}(\hat{\theta}) = \frac{I(\theta)}{\text{Var}(\hat{\theta})} \leq 1$$

- When $\text{efficiency}(\hat{\theta}) = 1$, then $\hat{\theta}$ is said to be an optimal estimator for θ (also known as the Minimum Variance Unbiased Estimator or MVUE).

Maximum Likelihood Estimators: properties

Under fairly weak assumptions, MLE estimators have several nice properties:

- **Equivariance**:
  - Let $\hat{\theta}$ be the MLE estimator for θ, and let $g(\cdot)$ be a bijective transform (one-to-one).
  - Let γ be a different parameterization such that $\gamma = g(\theta)$.
  - Then $\hat{\gamma} = g(\hat{\theta})$.
  - ⇒ We can easily find MLE for transforms of the parameter θ, e.g., useful when changing units, scale, or parametrization.
- **Consistency**: As the sample size increases, the MLE converges to the true parameter value.
- **Asymptotic efficiency**: Among all well-behaved estimators, the MLE has the smallest variance as $n \to \infty$.
- **Asymptotic Normality**:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I(\theta)^{-1})$$

- As the sample size increases, the MLE estimator distribution approaches a Gaussian centered around the true θ.
- Note: $I(\theta)$ is the Fisher information computed at the true value θ (often computed analytically).
- **"Often" a function of a sufficient statistic.**

BONUS: MLE song for «tuning parameters» Cringest moment in class

Examples of Maximum Likelihood: estimators for popular distributions

Exponential distribution: estimator for average number of events

Let X be a random variable that models the decay time of a radioactive nucleus, such that:

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

where λ is the average number of decays per year.

- What is the MLE for λ?
  - $f(X;\lambda) = \lambda e^{-\lambda X}$
  - Given a random sample $\boldsymbol{X_n} = \{X_1, \dots X_n\}$ of IID components, then we can write the likelihood as:
$$\mathcal{L}(\lambda; \boldsymbol{X_n}) = \prod_{i=1}^{n} f(\lambda; X_i) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i}$$
  - Taking the log and deriving with respect to λ this becomes:
$$\ell(\lambda) = \sum_{i=1}^{n} [\log(\lambda) - \lambda X_i] = nlog(\lambda) - \lambda \sum_{i=1}^{n} X_i \xrightarrow{\frac{d}{d\lambda}} S(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i$$
  - Finally, setting $S(\lambda) = 0$ and solving for λ:
$$S(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0 \longrightarrow \widehat{\lambda^{MLE}} = \frac{n}{\sum_{i=1}^{n} X_i} = \bar{X}^{-1}$$
- Is λ^{MLE} unbiased estimator for λ?
  - $E\left[\widehat{\lambda^{MLE}}\right] = E\left[\frac{n}{\sum_{i=1}^{n} X_i}\right] = nE\left[\frac{1}{\sum_{i=1}^{n} X_i}\right]$
  - It is possible to show that for $Y = \sum_{i=1}^{n} X_i$, where $X_i \sim Exp(\lambda)$, then $Y \sim Gamma(n, \lambda)$
  - Also, $Z = \frac{1}{Y} \sim Inv - Gamma(n, \lambda)$, for which we know that $E[Z] = \frac{\lambda}{n-1}$
  - Finally, $E\left[\widehat{\lambda^{MLE}}\right] = nE[Z] = \frac{n}{n-1}\lambda$
  - → **$\widehat{\lambda^{MLE}}$ is a biased estimator for λ!**
  - → However, all MLE asymptotic properties hold

Exponential distribution: estimator for average lifetime

Let X be a R.V. that models the decay time of a radioactive nucleus, such that:

$$X \sim \text{Exp}(\tau)$$

where τ is average lifetime in years.

- What is the MLE for τ?
  - $\tau = \lambda^{-1}$ is an invertible function of the parameter λ
  - We can leverage the equivariance property of MLEs: $g(\widehat{\lambda})^{MLE} = g\left(\widehat{\lambda^{MLE}}\right)$
  - Hence: $\widehat{\tau^{MLE}} = \left(\widehat{\lambda^{MLE}}\right)^{-1} = \bar{X}$

- Is τ^{MLE} unbiased estimator for τ?
  - By the Law of Large Numbers we know that the sample mean is an unbiased estimator for the population mean, i.e. $\mathrm{E}[\bar{X}] = \mu$
  - in this case μ is the average lifetime in the population (all radioactive nuclei of that type)

  → $\widehat{\tau^{MLE}}$ is unbiased estimator for τ

Gaussian distribution: estimator for μ

Let X be a R.V. that models the measurement of the speed of light in a given medium. Repeated measurements yield slightly different results that can be described by a Gaussian distribution, with unknown mean parameter μ:

$$X \sim N(\mu, \sigma^2)$$

where μ is the true measurement value and σ^2 (Nuisance parameter, we assume this is fixed and known a priori) is the instrument resolution

- What is the MLE for μ?

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

Nuisance parameter, we assume this is fixe and known a priori (e.g. instrument specs)

Then we can write the likelihood as:

$$\mathcal{L}(\mu; \boldsymbol{X_n}, \sigma^2) = \prod_{i=1}^{n} f(\mu; X_i) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{\wedge}\left\{-\frac{\sum_i^n\{(X_i-\mu)^2\}}{2\sigma^2}\right\}$$

Taking the log and deriving with respect to μ this becomes:

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n}(X_i-\mu)^2 \xrightarrow{\frac{d}{d\mu}} S(\mu) = \frac{\sum_i(X_i-\mu)}{\sigma^2}$$

Finally, setting $S(\mu) = 0$ and solving for μ:

$$\sum_i (X_i - \mu) = 0 \longrightarrow \widehat{\mu^{MLE}} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}$$

- For the Law of Large Numbers we already know μ^{MLE} is unbiased
  - Note: although we did not need σ^2 to compute μ^{MLE}, the standard errors of μ^{MLE} still depend on it

What if σ^2 is unknown and we want to estimate it?

- $\sigma^2_{MLE} = S^*$
- We already know this is a biased but consistent estimator for σ^2
  -> MLEs are not always unbiased for finite samples!

Poisson distribution: asymptotic distribution of λ^{MLE}

Let X be a R.V. that models the dark count of photons of our detector, i.e. the number of false positive counts our detector registers due to thermal noise and background effect. Since the detector works very well in general, we can assume dark counts are rare events, so the process can be described by a Poisson distribution, with unknown intensity parameter λ.

- What is the asymptotic distributions of λ^{MLE}?

  From MLE asymptotic properties we know that: $\widehat{\lambda^{MLE}} \rightarrow_{n\to\infty} N\left(\lambda, I^{-1}(\lambda)\right)$

  - We assume regularity conditions hold (that is generally the case)

  Hence, we just need to derive the Fisher information to express the variance of the asymptotic distribution

  - $I(\lambda) = -E\left[\frac{d^2}{d\lambda^2}\ell(\lambda)\right]$
  - $p(X;\lambda) = \frac{e^{-\lambda}\lambda^X}{X!}$; $\mathcal{L}(\lambda;X) = \prod_{i=1}^{n}\frac{e^{-\lambda}\lambda^{X_i}}{X_i!} = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n}X_i}}{\prod_{i=1}^{n}X_i}$; $\ell(\lambda;X) = -n\lambda + \log(\lambda)\sum_{i=1}^{n}X_i - \sum_{i=1}^{n}log(X_i)$
  - $S(\lambda) = \frac{d}{d\lambda}\ell(\lambda;X) = -n + \frac{1}{\lambda}\sum_{i=1}^{n}X_i$; $\frac{d^2}{d\lambda^2}\ell(\lambda) = -\frac{1}{\lambda^2}\sum_{i=1}^{n}X_i$

  Now we have general formulas, but in practice we do not know λ → estimate through $\widehat{\lambda^{MLE}}$

  - $\widehat{\lambda^{MLE}} = \bar{X}$; $\mathrm{I}(\lambda) = -E\left[\frac{d^2}{d\lambda^2}\ell(\lambda)\Big|_{\lambda=\widehat{\lambda^{MLE}}}\right] = -\mathrm{E}\left[-\frac{n\bar{X}}{\bar{X}^2}\right] = \frac{1}{\lambda}$

  Finally, we can approximate the asymptotic distribution by sample estimates: $\widehat{\lambda^{MLE}} \rightarrow_{n\to\infty} N(\bar{X}, \bar{X})$

1. Hypothesis testing

Outline

- How to test hypotheses?
- Hypothesis testing
- Examples

How to test hypothesis?

Hypothesis Testing:
A structured approach to evaluate claims about a population using sample data under uncertainty.
Competing Hypotheses:
- **Null Hypothesis** H_0: Represents the given assumption or status quo.
- **Alternative Hypothesis** H_a: Challenges the null hypothesis.

Determine whether data are compatible with the null hypothesis based on a figure of merit (e.g., test statistic, p-value).

Hypothesis Testing: Key Ingredients

Components:

1. **Null Hypothesis** H_0:
   - Hypothesis to test (e.g., "The mean temperature of a star is 5000K").
2. **Alternative Hypothesis** H_1:
   - Competing hypothesis (e.g., "The mean temperature of a star is not 5000K").
3. **Test Statistic:**
   - A figure of merit summarizing sample data to assess H_0.
4. **Significance Level (α):**
   - Confidence level (common choices: 0.05, 0.01; in physics: 5σ\sigma).
5. **p-value (p):**
   - Probability of observing the data (or something more extreme) under H_0.
6. **Decision Rule:**
   - If $p \leq \alpha$: Reject H_0.
   - If $p > \alpha$: Fail to reject H_0.

Hypothesis testing: intuition

Frequentist Approach:

"Assume H_0 is true and look for evidence in the sample that contradicts this assumption."

Analogy: Courtroom Trial

- H_0: The defendant is innocent.
- H_1: The defendant is guilty.
- If evidence is strong enough (low p-value), we reject H_0 (innocence).
- If evidence is weak (high p-value), we fail to reject H_0 (no conviction).

Hypothesis Testing: Formal Definition

Let X describe a random process with probability distribution $f(X,\theta)$, where θ is unknown.

1. **Null Hypothesis (H_0):**
   - Hypothesis about the value of θ, e.g., $\theta = \theta_0$.
   - Assumes that the true value of θ is a specific value θ_0.
   - This is the hypothesis we want to test.
2. **Alternative Hypothesis (H_1):**
   - Hypothesis about the parameter value that differs from H_0.
   - Examples:
     - Simple hypothesis: $H_1 : \theta = \theta_1$.
     - Composite hypothesis: $H_1 : \theta \neq \theta_0$ or $H_1 : \theta > \theta_0$.
3. **Data Representation (X):**
   - Let $X = \{X_1, \ldots, X_n\}$ represent n IID realizations of the random variable X.
   - Example: Observations of a single particle, event, or whole experiment.

Key Idea: Hypothesis testing determines a decision rule to evaluate whether observed data X are compatible with H_0.

Hypothesis Testing: Decision Rule

Decision Basis

- Ideally, base decisions on $P(H_0 \mid X)$, but this is not possible in a frequentist approach.
  - This is because the frequentist framework considers probabilities as long-run frequencies of events. In this approach, P(H0|X) (the probability of the null hypothesis given the data) does not have a meaningful interpretation because hypotheses are treated as fixed (true or false), not random variables.
- Instead, use the likelihood $P(X \mid H_0)$ and ask:
  - *"If H_0 is true, what is the probability of observing the sample data X?"*

Reason: If sample data are very unlikely under H_0, we question the validity of H_0.

In practice, we need two elements to apply this principle:

1. **Test Statistic ($t(X)$):**
   - that is a characteristic of sample data used as a benchmark.
   - Must be based on sample data.
   - We must know how to compute $P(t(X) \mid H_0)$.
2. **Significance Level (α):**
   - Minimum probability threshold we are willing to accept.
   - Outcomes rarer than α lead to rejection of H_0.
   - Also called the "size of the test."

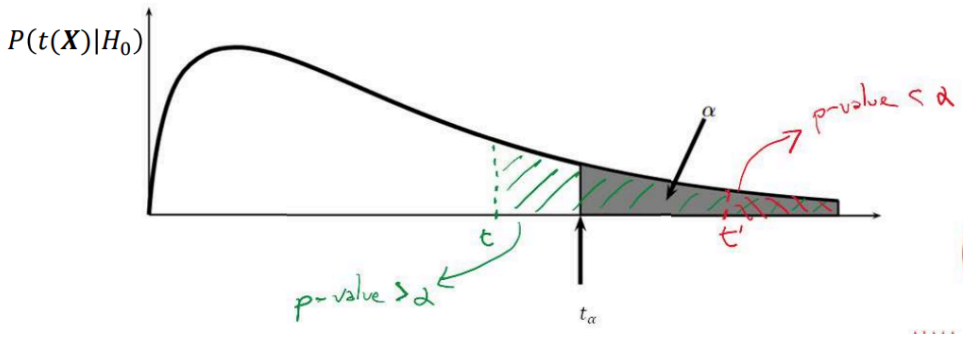
The key is to choose a test statistic for which we know $P(t(X) \mid H_0)$.
When that is the case, hypothesis testing consists in two steps:

1. **Compute the p-value:**
   - The probability of observing outcomes at least as rare as the sample data (i.e., equally or more unlikely).
2. **Compare the p-value with the significance level α:**
   - If p-value $\leq \alpha \rightarrow$ Reject H_0.
   - If p-value $> \alpha \rightarrow$ Fail to reject H_0.

Decision Rule: If the observed data is rarer than the significance level α, reject the null hypothesis H_0.

Interpretation: Statistically significant evidence against H_0.

Visual Representation



Here 2 examples, a simple one and a more complex to use also the $t(x)$.

Example 1

Hypothesis testing: example

Imagine you are playing heads or tails tossing a fair coin with a friend. You choose to always bet on heads, but after 10 trials you only won once, so you start wondering whether the coin is actually fair.

How can you check your doubts? → hypothesis testing can help!

- How can we model the coin flip trials? → Binomial : $\boldsymbol{X \sim Bin(n, p)}$
- What is our null hypothesis? $\boldsymbol{H_0: p = 0.5}$
- And the alternative? $\boldsymbol{H_1: p > 0.5}$
- How do we compute the p-value?
  - The p-value is the probability of observing outcomes at least as rare as the sample data
  - In our case, under the null hypothesis (coin is fair), the only other outcome at least as extreme as having 9 tails in 10 flips is having 10 tails
  - $p - value = P(X = 9|H_0) + P(X = 10|H_0) = P(X = 9|p = 0.5) + P(X = 10|p = 0.5)$
    $= \binom{10}{9} 0.5 \cdot 0.5^9 + \binom{10}{10} 0.5^{10} = 0.0107$
  - What is the significance level? This is something you can decide
    - One one side, you do not want to lose money, so do not set this too low → otherwise, in case H_0 is false you will lose a lot of money before collecting enough evidence against it
    - On the other: you do not want to argue with your friend, so do not set this too high → otherwise, you might reject H_0 due to random fluctuations even if it is true
    - $\alpha = 0.05$ is commonly used, i.e. we reject outcomes rarest than 5% under H_0
- Decision? $p - value < \alpha$ → reject H_0: the coin is unlikely to be fair based on observed data

Note:

Although arbitrary, the **significance level must be set before looking at the data/p-value!**

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Example 2

Physics Example: Particle Lifetime Hypothesis Testing

We are testing whether a particle decays with a mean lifetime of$\tau_0 = 2.5\,\mu s$ based on experimental data.

Hypotheses

1. **Null Hypothesis (H_0):** The particle's mean lifetime is $\tau_0 = 2.5\,\mu s$
2. **Alternative Hypothesis (H_1):** The particle's mean lifetime is different from $\tau \neq 2.5\,\mu s$

This is a **two-tailed test** because we are testing for deviation in either direction.

Data Representation

- Observed decay times of $n = 30$ particles are recorded.
- Sample mean decay time: $\bar{t} = 2.3\,\mu s$

The decay times are assumed to follow an **exponential distribution,** which leads to the sample mean $\bar{t}$ being normally distributed for large n:

$$\bar{t} \sim \mathcal{N}(\mu = \tau_0, \sigma = \frac{\tau_0}{\sqrt{n}})$$

where:

- $\mu = \tau_0 = 2.5\,\mu s$
- $\sigma = \frac{\tau_0}{\sqrt{n}} = \frac{2.5}{\sqrt{30}} \approx 0.457\,\mu s$

Significance Level

We choose a **significance level** of $\alpha = 0.05$, meaning we will reject H0 if the probability of observing such an extreme result (or more extreme) is less than 5%.

The test statistic is the **sample mean**:

$$t(X) = \bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

We compute the probability of observing a mean decay time at least as extreme as the observed value $\bar{t} = 2.3\,\mu s$, under H0.

The **p-value** is:

$$p = P(\bar{t} \leq 2.3\,\mu s \text{ or } \bar{t} \geq 2.7\,\mu s)$$

since the normal distribution is symmetric

$$P(\bar{t} \leq 2.3) = \int_{-\infty}^{2.3} f(t)dt$$

where f(t) is the PDF of $\mathcal{N}(2.5, 0.457)$. This is done using the cumulative distribution function (CDF).

The **p-value** is $p = 0.661$.
Conclusion: Since $p > \alpha$, we **fail to reject H_0**.

Interpretation

There is not enough evidence to conclude that the particle's mean lifetime differs from $\tau_0 = 2.5\,\mu s$.
The observed data is consistent with the null hypothesis.

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

Here instead a second way to look at hypotesis testing.

2. Hypothesis Testing: Rejection Regions

Another way to interpret hypothesis testing is by defining a **critical region** or **rejection region** RR_α:

- The rejection region includes rare outcomes under H_0.
- The probability of the critical region is the significance level α, i.e., $P(X \in RR_\alpha \mid H_0) \leq \alpha$.

Decision Rule

- If the observed outcome belongs to the rejection region, **reject** H_0.

The rejection region can be expressed in terms of a **critical value** k:

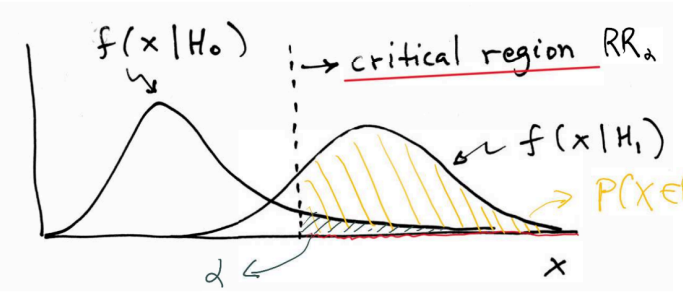
$$P(X \in RR_\alpha \mid H_0) \leq \alpha \implies P(X > k \mid H_0) \leq \alpha$$

- k separates the rejection region from the rest of the sample space.
- It can be $X > k$ or $X < k$ depending on the test.

To choose an appropriate rejection region, consider H_1:

- Place RR_α where outcomes are **rare under** H_0 **but common under** H_1
- Example:
  - $P(X \in RR_\alpha \mid H_0)$ is low, but $P(X \in RR_\alpha \mid H_1)$ is high.

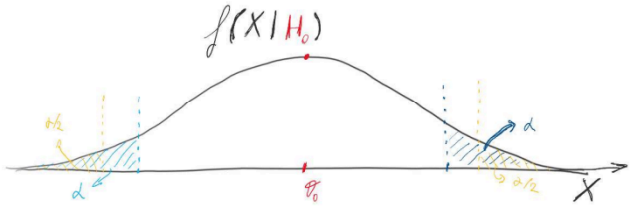
This framework ensures we reject H_0 only when there's sufficient evidence favoring H_1.



Example:

Imagine you are testing a simple null hypothesis, $H_0: \theta = \theta_0$. The probability distribution of the data under H_0 is reported in the plot below. Graphically indicate RR_α for each of the following alternative hypothesis scenarios:

- $H_1: \theta = \theta_1 > \theta_0$
  - This suggests rare outcomes under H_0 but "common" under H_1 are values in the **right tail**
- $H_2: \theta < \theta_0$
  - Just the opposite situation → **left tail**



- And for $H_3: \theta \neq \theta_0$?
  - RR_α includes both tail, halving its size at each side to ensure a global significance level of α

→ H_1 and H_2 are said **one-sided** alternatives, while H_3 is called **two-sided**

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Interpreting test results

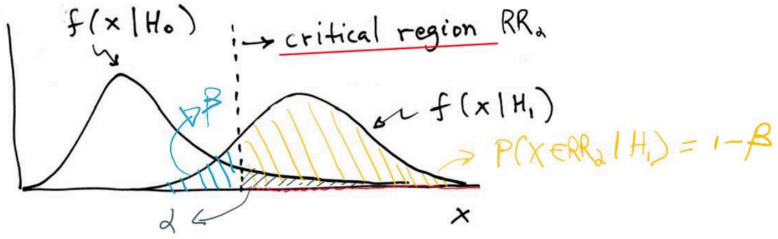
Regarding the final decision of a test, there are some subtle nuances to bear in mind

- **Rejecting** H_0:
  - Does not confirm H_0 is false or H_1 is true.
  - Indicates sufficient evidence against H_0 and in favor of H_1.
- **Failing to reject** H_0:
  - When p-value $> \alpha$:
    - Either H_0 is true.
    - Or H_0 is false, but the test has low power.
- **Misinterpretation of** p-value:
  - p-value is not the probability of the hypothesis: $p\text{-value} \neq P(H_0|X)$
  - Represents the probability of observed data under H_0: $P(X|H_0)$
  - Informally, it measures evidence against H_0.
  - Can be seen as the smallest significance level at which H_0 is rejected.
- **Scientific relevance**:
  - Statistically significant results may lack practical significance.
  - Example: $\theta \neq \theta_0$ but with negligible impact on the theory.

Type-I, Type-II errors

- **Type-I errors**: Reject H_0 when it is actually true.
  - Happens with probability $P(X \in RR_\alpha|H_0) \leq \alpha$.
  - Called the significance level or size of the test.
  - **Interpretation of** α: Probability of erroneously rejecting H_0.

- **Type-II errors**: Fail to reject H_0 when H_1 is true.
  - Happens with probability $P(X \in \Omega \setminus RR_\alpha | H_1) = \beta$.
  - $1 - \beta$ is called the power of the test.
  - **Interpretation of $1 - \beta$**: Probability of correctly rejecting a false null hypothesis when H_1 is true.



Uniformly Most Powerful test (UMP)

A test is said to be **Uniformly Most Powerful (UMP)** if:

- It maximizes the power for all possible values of the alternative hypothesis.
- For a fixed significance level α.
- In other words, no matter the true parameter value under H_1, the UMP test gives the best chance of rejecting H_0.

This is equivalent to requiring a single critical region to ensure maximum power independently of the alternative hypothesis, leading to a **model-independent test**.

- In High-Energy Physics (HEP), we often construct tests such that:
  - H_0: Standard Model (or "background only").
    - α: Probability of rejecting H_0 when it is true (false discovery rate, typically 5σ).
  - H_1: An interesting alternative theory (e.g., SUSY, Z', etc.).
    - We aim for high power with respect to any possible alternative new theory.
- Unfortunately, there is no general guarantee of having a model-independent test.
  - **Solution**: Select a critical region that maximizes the power for a specific H_1.

Intuition

In brief, UMP tests can be summarized as follows:

- **Power of a test**:
  - The probability of rejecting H_0 when H_1 is true.
  - Ideally, we want the power to be as large as possible because this means the test is sensitive to detecting true differences.
- **Why is UMP desirable?**:

    - A UMP test guarantees that, regardless of the true value of the parameter under H_1, the test will have the highest chance of detecting the alternative hypothesis.
- **Competing tests:**
    - Different tests can have different power properties.
    - A test might be powerful for one value of H_1 but not for others.
    - A UMP test, if it exists, ensures that no matter what value H_1 takes, it is the most powerful option.

How to find an UMP test?

Neyman-Pearson Lemma

The Neyman-Pearson lemma provides an elegant way to find a UMP test for testing simple hypotheses:

- Let H_0, H_1 be two simple hypotheses, i.e.,

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta = \theta_1 \neq \theta_0$$

- Let $t(X) = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)}$ be a test statistic defined as the ratio between the likelihoods under the two hypotheses, $\mathcal{L}(\theta \mid X, H_1)$ and $\mathcal{L}(\theta \mid X, H_0)$ respectively.
- For a fixed significance level α, the Neyman-Pearson lemma shows that:
  The decision rule: "reject H_0 when $t(X) > k$" is the UMP test, where k is chosen such that $P_{\theta_0}(t(X) > k) = \alpha$.

Limitations:

- This result relies on the assumptions on the underlying data distribution:
    - What if $\mathcal{L}(\theta)$ is incorrect, or if we do not know it?
- Even if we know $\mathcal{L}(\theta)$, deriving the distribution of the ratio is not always possible:
    - Requires numerical approximations.
- UMP tests do not always exist:
    - Neyman-Pearson lemma only holds for simple hypotheses.
    - What about composite hypotheses? (i.e., one-sided or two-sided tests)

Walt test

Let θ be a scalar parameter and $\hat{\theta}$ be its estimate. Also, let $\hat{\sigma}$ be the estimated standard error of $\hat{\theta}$.

Consider testing: $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

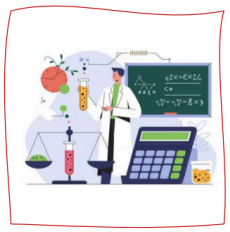
- When $\hat{\theta}$ is Normal:

$$\hat{\theta} \sim N(\theta_0, \sigma^2) \rightarrow \frac{\hat{\theta} - \theta_0}{\hat{\sigma}} \sim N(0, 1)$$

- Then we can define the Wald test as:
  - Choose test statistic $W = \frac{\hat{\theta}-\theta_0}{\hat{\sigma}}$
  - Choose critical value *k* based on asymptotic normal distribution
    - $P_{\theta_0}(|W| > k) = \alpha \rightarrow P_{\theta_0}(|Z| > k) = \alpha \rightarrow \begin{cases} P_{\theta_0}(Z > k) = \frac{\alpha}{2}, & \text{when } W > 0 \\ P_{\theta_0}(Z < -k) = \frac{\alpha}{2}, & \text{when } W < 0 \end{cases}$
    - Hence, we set *k* as the quantiles of a standard normal distribution $z_{\alpha/2}$, i.e. threshold for which $P_{\theta_0}(Z > z_{\alpha/2}) = \frac{\alpha}{2}$
    - In other words: the critical point is the value for which the cdf of a standard normal returns $1 - \frac{\alpha}{2}$, i.e. $z_{\alpha/2} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$
  - Equivalently: $\text{p} - \text{value} = P_{\theta_0}(|W| > |w|) = P_{\theta_0}(|Z| > |w|) = 2\Phi(-|w|)$
    where *w* is the value of *W* observed in the sample
- Decision rule:
  - Based on rejection region: Reject H_0 if $W > z_{\alpha/2}$ or $W < -z_{\alpha/2}$
  - Based on p-value: Reject H_0 if $p - value < \alpha$

Example:

Let X be a random variable describing the average surface temperature of a star, such that $f(X, \theta)$ is its probability distribution and θ represents the true mean temperature.
Imagine we observe a sample $\boldsymbol{X} = \{X_1, \dots, X_n\}$ of $n = 30$ experimental measurements of the surface temperature, and we get that $\bar{x} = 5008K$ and $s^2 = 64K^2$.
Now, our current theory postulates that $\theta = 5000K$. Can we claim that our experiment is a new discovery?



- **Null hypothesis (H_0)**: the hypothesis we want to test is $\text{H}_0: \theta = 5000K$
- **Alternative hypothesis (H_1)**: $\text{H}_1: \theta \neq 5000K$
- **Test Statistic**: $t(\boldsymbol{X}) = W = \frac{\hat{\theta}-\theta_0}{\hat{\sigma}}$
  - We need to estimate $\hat{\theta}$ and $\hat{\sigma}$ → we can use $\bar{x}$ and $\sqrt{\{s^2\}}$, respectively
  - By LLN we know that $W = \frac{\bar{X}-\theta_0}{s/\sqrt{n}} \sim_{n\to\infty} N(0,1)$
- **Significance Level (α)**: $\alpha = 5\sigma \approx 2.9 \times 10^{-7}$
- **p-value (p)**:
  - How to compute this? Given the asymptotic distribution:

$$\text{p} - \text{value} = P_{\theta_0}(|W| > |w|) = P_{\theta_0}\left(|Z| > \left|\frac{\bar{x} - \theta_0}{\frac{s}{\sqrt{n}}}\right|\right) = 2\Phi\left(-\left|\frac{5008 - 5000}{\frac{8}{\sqrt{30}}}\right|\right) \approx 2\Phi(-5{,}47) \approx 4.5 \times 10^{-8}$$

- **p-value (p)**: $P(X \in RR_\alpha | \theta = 5000K) \approx 4.5 \times 10^{-8}$
- **Decision Rule**:
  - If p-value ≤ α → Reject H_0 → **new discovery!**

Comparing two sample proportions

We are studying the decay rates of two radioactive isotopes, A and B, over a fixed time period t. Let X be the random variable describing these processes, such that: $X_A \sim Bern(p_A)$, $X_B \sim Bern(p_B)$, and $X_A \perp X_B$.

Now imagine we observe two sufficiently large samples of decays for each isotope, i.e., $n_A, n_B > 30$, and we want to test whether the decay probabilities of the two isotopes are significantly different.

For this problem, we can use a Wald test (Z-test) for comparison of two sample proportions (frequencies).

- **Hypothesis:** $H_0 : p_A = p_B$ vs $H_1 : p_A \neq p_B$
- **Test statistic:** $Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}} \xrightarrow{n_A+n_B\to\infty} N(0,1)$

$\rightarrow$ p-value and rejection region can be computed based on the standard normal distribution.

Note:
When the sample size is small, we can use the exact Fisher test

---

Comparing two sample means

We are studying the thermal conductivity of two materials. Let X be the random variable describing these processes, such that $X_A \perp X_B$.

Now imagine we observe two sufficiently large samples, and we want to test whether the average insulation properties of the two materials are significantly different.

For this problem, we can use a Wald test (Z-test) for comparison of two sample means.

- **Hypothesis:** $H_0 : \mu_A = \mu_B$ vs $H_1 : \mu_A \neq \mu_B$
- **Test statistic:** $Z = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \xrightarrow{n_A+n_B\to\infty} N(0,1)$
    - If σ_A^2, σ_B^2 are unknown, we estimate them through their sample counterparts s_A^2, s_B^2.

$\rightarrow$ p-value and rejection region can be computed based on the standard normal distribution.

Note:
When the sample size is small, if $X \sim N$ and homoscedastic (same variance), we can use the t-Student test.

Comparing two sample variances

We are evaluating two different particle detector designs for a new high-energy physics experiment. Let X be the random variable describing measured energy of each detector for a known calibration source, such that: $X_A \sim N(\mu_A, \sigma_A^2)$, $X_B \sim N(\mu_B, \sigma_B^2)$, and $X_A \perp X_B$.

Now imagine we observe two sufficiently large samples, and we want to determine if there is a significant difference in detector precision when measuring particle energies.

For this problem, we can use a Fisher-Snedecor test (F-test) for comparison of two sample variances:

- **Hypothesis:** $H_0 : \sigma_A^2 = \sigma_B^2$ (homoscedastic) vs $H_1 : \sigma_A^2 \neq \sigma_B^2$
- **Test statistic:** $F = \frac{s_A^2}{s_B^2} \overset{H_0}{\longrightarrow} F(n_A - 1, n_B - 1)$
  - where s_A^2, s_B^2 are the sample variances, and F is the Fisher distribution

$\rightarrow$ p-value and rejection region can be computed based on the Fisher-Snedecor distribution.

Comparing more sample means

What if we want to compare more than two sample means? For example, compare K samples and test whether they all belong to the same population.

If we assume the measurements in each sample are:

- Independent
- Normal
- Homoscedastic, i.e., they have the same variance

Then we can use the Analysis Of Variance test (ANOVA-test) for comparison of K sample means:

- **Hypothesis:**
  $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K = \mu \ \ \backslash \ \ \text{vs} \ \ \backslash \ \ H_1 : \mu_i \neq \mu_j$ for at least a couple $i, j_{i,j=1,\ldots,n}$
- **Test statistic:**
  $F = \frac{V(X)_{\text{between}}}{V(X)_{\text{within}}} \overset{H_0}{\longrightarrow} F(K-1, n-K)$, where:
  - $V(X)_{\text{between}} = \frac{\sum_{k=1}^{K} n_k (\bar{X}_k - \bar{X})^2}{K-1}$
  - $V(X)_{\text{within}} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} (X_{k,i} - \bar{X}_k)^2}{n-K}$
  - $n = \sum_{k=1}^{K} n_k$

$\rightarrow$ p-value and rejection region can be computed based on the Fisher-Snedecor distribution.

Testing binned distribution

Suppose we are analyzing data from a particle physics experiment where we've measured the invariant mass of a large number of particle decay events. We want to

determine if our observed mass distribution fits a theoretical model, which could confirm or refute the presence of a new particle.

To do so, we can organize mass measurements into a histogram with H bins and compare the entries in each bin to the theoretical distribution of a given model. Let X_h be the random variable describing the number of events in the i-th bin, such that $X_h \perp X_k$ for $h \neq k$.

If we have a sufficiently large sample in each bin (say > 5), then we can use a Pearson Chi-squared test (χ^2-test):

- **Hypothesis:** $H_0 : p_h = \pi_h \, \forall \, h = 1, \ldots, H$ vs $H_1 : p_h \neq \pi_h$ for at least one $h_{h=1,\ldots,H}$
- **Test statistic:** $\chi^2 = \sum_{h=1}^{H} \frac{(n_h - n\pi_h)^2}{n\pi_h} \overset{H_0}{\longrightarrow} \chi^2(H - M - 1)$
  - where n_h is the observed number of events in bin h, and $n\pi_h$ is the expected one.
  - M is the number of fitted parameters.

$\rightarrow$ p-value and rejection region can be computed based on the Chi-squared distribution.

Goodness of Fit

Goodness of fit refers to how well a statistical model fits a set of observations.

- It describes the discrepancy between observed values and expected values under the model in question.
- Can be seen as a particular case of hypothesis testing:
  - More general alternative $\rightarrow H_1$: all possible alternatives.
- **Note:** Often we test for goodness of fit, but our hope is poor agreement:
  - A failed test means rejecting H_1: current knowledge $\rightarrow$ discovery!

Pearson's χ^2 Statistic

Test statistic for comparing observed data $\vec{n} = (n_1, \ldots, n_N)$ (n_i independent) to predicted mean values $\vec{\nu} = (\nu_1, \ldots, \nu_N)$:

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\sigma_i^2}, \quad \text{where } \sigma_i^2 = V[n_i]$$

(Pearson's χ^2 statistic)

χ^2 is the sum of squares of the deviations of the ith measurement from the ith prediction, using σ_i as the "yardstick" for the comparison.

For $n_i \sim \text{Poisson}(\nu_i)$, we have $V[n_i] = \nu_i$, so this becomes:

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}$$

Pearson's χ^2 Test

If n_i are Gaussian with mean ν_i and std. dev. σ_i, i.e., $n_i \sim N(\nu_i, \sigma_i^2)$, then Pearson's χ^2 will follow the χ^2 pdf (here for $\chi^2 = z$):

$$f_{\chi^2}(z; N) = \frac{1}{2^{N/2}\Gamma(N/2)} z^{N/2-1} e^{-z/2}$$

If the n_i are Poisson with $\nu_i \gg 1$ (in practice OK for $\nu_i > 5$), then the Poisson distribution becomes Gaussian, and therefore Pearson's χ^2 statistic here also follows the χ^2 pdf.

The χ^2 value obtained from the data then gives the p-value:

$$p = \int_{\chi^2}^{\infty} f_{\chi^2}(z; N)\, dz$$

The χ^2 per Degree of Freedom

Recall that for the chi-square pdf for N degrees of freedom:

$$E[z] = N, V[z] = 2N. E[z] = N, \quad V[z] = 2N$$

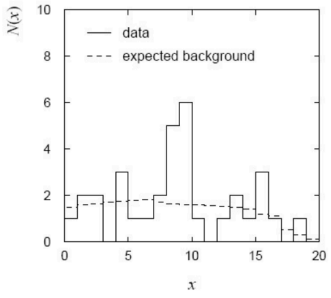
This makes sense: if the hypothesized ν_i are right, the RMS deviation of n_i from ν_i is σ_i, so each term in the sum contributes ~ 1.
One often sees χ^2/N reported as a measure of goodness-of-fit. But it is better to give χ^2 and N separately. Consider, e.g.:

$$\chi^2 = 15,\ N = 10 \ \rightarrow\ p\text{-value} = 0.13,$$

$$\chi^2 = 150,\ N = 100 \ \rightarrow\ p\text{-value} = 9.0 \times 10^{-4}$$

i.e., for N large, even a χ^2 per dof only a bit greater than one can imply a small p-value, i.e., poor goodness-of-fit.

Example:



← This gives

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} = 29.8$$

for $N = 20$ dof.

Now need to find p-value, but... many bins have few (or no) entries, so here we do not expect χ^2 to follow the chi-square pdf.
→ Remember to check $n_i > 5$

Likelihood Ratio Test (LRT)

The Look-Elsewhere Effect

Definition

- The **look-elsewhere effect**, also known as **multiple testing**, refers to the increased probability of a false positive result (Type I error) when multiple independent tests are performed on the same dataset.
- **Intuition:** Repeatedly looking for deviations from known distributions increases the chance of observing false positives due to random fluctuations.

Example

- Suppose a model for a mass distribution predicts a peak at mass m with amplitude μ. The observed data show a bump at mass m_0.
- **Question:** How consistent is this bump with the no-bump hypothesis ($\mu = 0$)?

Hypothesis Testing

1. **If m_0 is known a priori:**
   - Compute a local p-value for the specific mass m_0.
2. **If m_0 is not fixed:**
   - Compute a global p-value allowing m_0 to move freely:

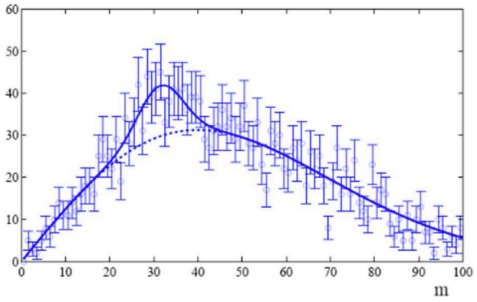
$$\alpha_{\text{global}} \approx \alpha_{\text{local}} \times N$$

   where N is the number of independent tests.

Correction Methods

- **Monte Carlo simulations:** Simulate the full testing process to account for multiple testing.
- **Bonferroni corrections:** Adjust α by dividing it by the number of tests.
- **Benjamini-Hochberg procedure:** Control the false discovery rate (FDR).

The graph shows data with a bump around m_0, which could either represent a true signal or a fluctuation. Proper statistical methods are needed to determine significance.



The significance of an observed signal

- Total events consist of:
  - n_b: Events from known processes (background).
  - n_s: Events from a new process (signal).
- If n_s and n_b are Poisson random variables with means s and b, then $n = n_s + n_b$ is also Poisson with mean $s + b$.

Probability Distribution:

$$P(n; s, b) = \frac{(s+b)^n e^{-(s+b)}}{n!}$$

Example: Observing $n_{\text{obs}} = 5$

- Background mean $b = 0.5$. Should we claim evidence for a new discovery?

Hypothesis $s = 0$ (no signal):

$$p\text{-value} = P(n \geq 5; b = 0.5, s = 0) = 1.7 \times 10^{-4} \neq P(s = 0)!$$

Significance from p-value

- **Significance Z:** The number of standard deviations a Gaussian variable would need to fluctuate in one direction to give the same p-value.

Relation Between p and Z:

$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \Phi(Z)$$

$$Z = \Phi^{-1}(1 - p)$$

- Φ: Cumulative distribution function of the standard normal distribution.

The Significance of a Peak

Hypothesis Testing

- Each bin (observed) is a Poisson random variable.
- Means are given by dashed lines (background).

Example:

- In the two bins with a peak:
  - Observed entries: $n = 11$.
  - Background mean: $b = 3.2$.

p-value for $s = 0$:

$$P(n \geq 11; b = 3.2, s = 0) = 5.0 \times 10^{-4}$$

Conclusion: The small p-value indicates the peak is unlikely under the background-only hypothesis.

Questions to Consider:

- **Look-Elsewhere Effect (LEE):**
  - How many x distributions have been analyzed?
    - For example, looking at 1000 histograms increases the probability of finding a 10^{-3} effect.
  - Adjust for the probability of finding a peak anywhere in the histogram.

- **Resolution Consistency:**
  - Is the observed width consistent with the expected x resolution?
  - Take an x window several times the resolution for verification.
- **Analysis Cuts:**
  - Were the cuts adjusted to enhance the peak? If so, freeze the cuts and repeat the analysis with new data.
- **Decision to Publish:**
  - Evaluate whether the observed effect is robust enough to justify publication.

When to Publish: Why 5 Sigma?

- **HEP Standard:** A p-value of 2.9×10^{-7}, corresponding to a significance $Z = 5$ (5-sigma), is typically required to claim a discovery.

Reasons for a High Threshold:

1. **Cost of False Discovery:**
   - Announcing a false discovery has significant consequences.
2. **Uncertainties in the Model:**
   - Address systematic and statistical uncertainties.
3. **Look-Elsewhere Effect:**
   - Correct for multiple testing.
4. **Extraordinary Claims:**
   - "Extraordinary claims require extraordinary evidence." – Carl Sagan

Key Reminder: The p-value is the **first step**, not the sole criterion for publishing. Consider how compatible the data are with the new phenomenon.

- The p-value quantifies the probability that the background-only model explains the observed fluctuation.
- **Not intended to address:**
  - Hidden systematics or high thresholds for a significant discovery.

Adjusted Threshold:

- If LEE is well-managed, the threshold for discovery could reasonably be closer to 3σ than 5σ.

Combining p-values

- **Scenario:** Two experiments test the same hypothesis H_0:
  - Experiment 1 reports 3σ, Experiment 2 reports 5σ.
  - How to combine the p-values p_1 and p_2?

Challenges:

- **Wrong Approach:** $p_{\text{comb}} = p_1/p_2$.

- **Correct Approach:** Use the Fisher method:

$$p_{\text{comb}} = P(p_1 p_2 [1 - \ln(p_1 p_2)])$$

  - This ensures the combined p-value reflects the joint significance.

Notes:

- The Fisher method generalizes to multiple p-values but is not associative.
- A combined test statistic t_{comb} should be computed when possible.

Interval estimation

Outline

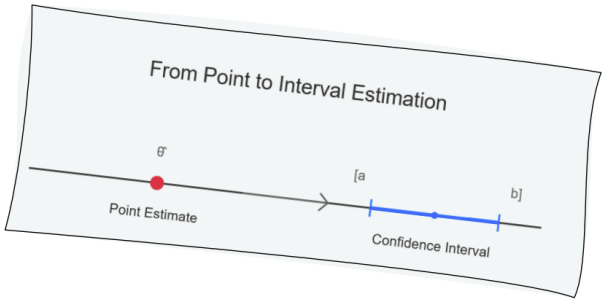
- How to measure uncertainty about estimates?
- Interval estimation
- Examples

How to measure uncertainty about parameter estimates?

From Point Estimates to Intervals

What We Have Covered So Far:

1. **Methods for Point Estimation:**
    - Maximum Likelihood Estimation (MLE).
    - Method of Moments (MoM).
2. **Point Estimators' Properties:**
    - Bias
    - Consistency
    - Efficiency
    - Sufficiency
3. **Sampling Distributions:**



- **BUT:** Is a single number (point estimate) enough?

Example: measuring the Higgs boson mass

- Point estimate: m_H = 125.35 GeV
- How confident are we about that specific value?
- What other values are also plausible
- What values can we rule out instead?

.

The need for interval estimates

How can we improve and complete the information of point estimates?
Point estimates are affected by various uncertainties:

- **Statistical fluctuations**
- **Finite sample size**
- **Measurement precision**
- **Systematic effects**

A **range of values** (interval estimates) is safer and more informative than a single-point estimate.

Interval Estimation

- **Goal:** Rigorously quantify uncertainty.
- Builds on the estimator's sampling distribution.
- Key aspects:
  - $\hat{\theta}$ is random; θ is fixed.
  - Shape of the distribution depends on:
    - Sample size
    - True parameter value
    - Estimation method

Confidence Intervals

A confidence interval $[a, b]$ provides a range of plausible values for a parameter with a given confidence level.

Key Idea

- Balance between:
  - **Width (precision):** Narrow intervals give precise estimates.
  - **Confidence (reliability):** Higher confidence increases reliability but widens the interval.

Simple Approach

Let $\hat{\theta} \sim g(\hat{\theta}; \theta)$ be an estimator with PDF $g(\hat{\theta}; \theta)$.

- g Refers to the **probability density function (PDF)** of the estimator $\hat{\theta}$ for a parameter θ. This PDF describes the distribution of the estimator $\hat{\theta}$ under repeated sampling, given the true value of the parameter θ.
- Provide uncertainty as:

$$\hat{\theta}_{\text{obs}} \pm \hat{\sigma}_{\hat{\theta}}$$

- $\hat{\theta}_{\text{obs}}$: Observed value of the estimator.
- $\hat{\sigma}_{\hat{\theta}}$: Sample estimate of the standard deviation (standard error) of $g(\hat{\theta};\theta)$.

- Typically used for error bars in plots
- **Special Case: Gaussian** $g(\hat{\theta};\theta)$
  - Confidence can be quantified precisely.
  - **Note:** This assumption does not always hold.

Formal Definition

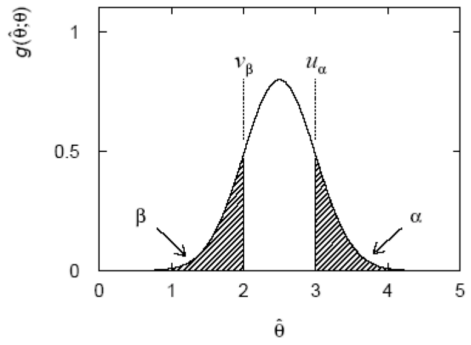
Let $v_\beta(\theta)$ and $u_\alpha(\theta)$ be the lower and upper bounds of an interval for $\hat{\theta}$. The confidence level is $1-\alpha-\beta$.

- **Intuition:** Find endpoints $[v_\beta(\theta), u_\alpha(\theta)]$ such that:

$$P\left(v_\beta(\theta) \leq \hat{\theta} \leq u_\alpha(\theta)\right) = 1-\alpha-\beta$$

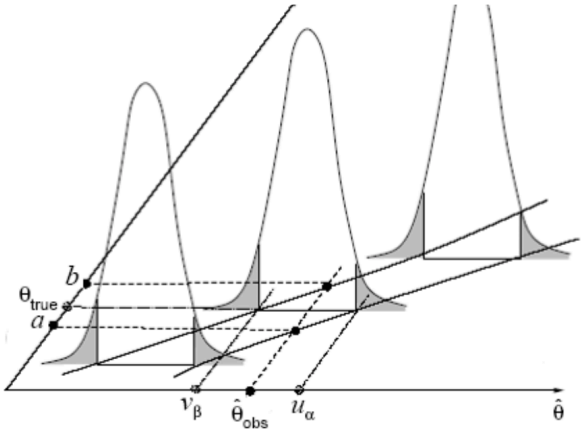
More Formally:

1. Define α and β:
   - $\alpha = P(\hat{\theta} \geq u_\alpha(\theta)) = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)d\hat{\theta}$
   - $\beta = P(\hat{\theta} \leq v_\beta(\theta)) = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta)d\hat{\theta}$
2. Solve these integrals for $v_\beta(\theta)$ and $u_\alpha(\theta)$:
   - By construction $[v_\beta(\theta), u_\alpha(\theta)]$ has $1-\alpha-\beta$ coverage for $\hat{\theta}$.



What about θ_{true}?

- When the estimator is well-behaved, the endpoints $v\beta(\theta)$, $u\alpha(\theta)$ are monotonic functions of θ
- If θ_{obs} falls in $\nu_\beta(\theta)$, $u_\alpha(\theta)$ then the interval (a,b) cover θ_{true}.



.

.

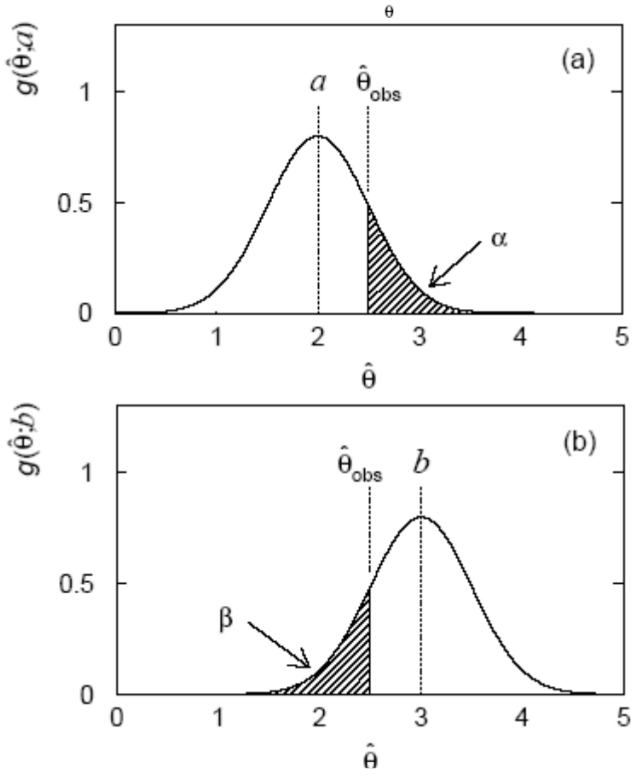
In practice

In practice, the recipe to find the interval (a, b) boils down to solving

- $\alpha = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)d\hat{\theta} = \int_{\hat{\theta}_{obs}}^{\infty} g(\hat{\theta};a)d\hat{\theta}$
- $\beta = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta)d\hat{\theta} = \int_{-\infty}^{\hat{\theta}_{obs}} g(\hat{\theta};b)d\hat{\theta}$

- a is the hypothetical value of θ such that $P(\hat{\theta} > \hat{\theta}_{obs}) = \alpha$
- b is the hypothetical value of θ such that $P(\hat{\theta} < \hat{\theta}_{obs}) = \beta$

Interpretation: if we were to repeat the experiment under same conditions many times, an interval built in this way would contain the true parameter value, θ, $(1 - \alpha - \beta) \cdot 100\%$ of the times



General Remarks on Confidence Intervals

1.
   - Is often reported as: $\hat{\theta}^{+d}_{-c}$, where $c = \hat{\theta} - a$ and $d = b - \hat{\theta}$
     - Exercise: what does $80.25^{+0.31}_{-0.25}$ mean?
       - 80.25 is our best estimate, $\hat{\theta}$
       - The interval is $[a = \hat{\theta} - c, b = \hat{\theta} + d] = [80, 80.56]$
       - If we repeat the experiment many times with same sample size and always construct the interval with this method, then the C.I. will contain the θ_{true} in $1 - \alpha - \beta$ fraction of the experiments
       - **Note: it doesn't mean $P(80 < \theta < 80.56) = 1 - \alpha - \beta$!!**

2.
   - We often use central confidence intervals, i.e. take $\alpha = \beta = \frac{\gamma}{2}$ → coverage $1 - \gamma$
     - Note: «central» does not mean symmetric about $\hat{\theta}$, only $\alpha = \beta$

3.
   - Sometimes we may be interested only in one-sided confidence intervals, i.e.
     - Set a as a lower limit such that $\mathrm{P}(a \geq \theta) = 1 - \alpha$ (here a is random, henceforth the use of probability)
     - Set b as a upper limit such that $\mathrm{P}(b \leq \theta) = 1 - \beta$ (same as above)

Examples

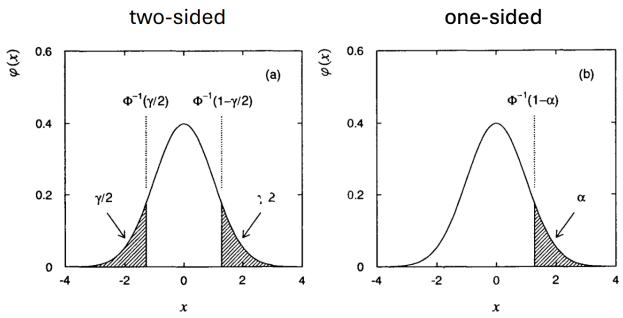
Mean of Gaussian confidence interval

When the distribution of the estimator, $\hat{\theta}$, is Gaussian, then the construction is much simpler. When the standard deviation $\sigma_{\hat{\theta}}$ is known, in fact:

- $\alpha = 1 - G(\hat{\theta}_{obs}; a, \sigma_{\hat{\theta}}) = 1 - \Phi\left(\frac{\hat{\theta}_{obs} - a}{\sigma_{\hat{\theta}}}\right)$
- $\beta = G(\hat{\theta}_{obs}; b, \sigma_{\hat{\theta}}) = \Phi\left(\frac{\hat{\theta}_{obs} - b}{\sigma_{\hat{\theta}}}\right)$

Which in turn implies that:

- $a = \hat{\theta}_{obs} - \sigma_{\hat{\theta}}\Phi^{-1}(1 - \alpha)$
- $b = \hat{\theta}_{obs} + \sigma_{\hat{\theta}}\Phi^{-1}(1 - \beta)$



- **Notes:**
  - In this case, the interval $[a, b] = [\hat{\theta}_{obs} - \sigma_{\hat{\theta}}, \hat{\theta}_{obs} + \sigma_{\hat{\theta}}]$ is the 68.3% confidence interval
  - This situation hold asymptotically for MLEs → just need to retrieve $\hat{\theta}_{MLE}$ and $\sigma_{\hat{\theta}}$ depending on data distribution
    - E.g. if $X \sim Binomial(n, p)$, then $\hat{p} = \bar{X}$ and $\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ → C.I. = $\hat{p} \pm z_{\frac{\alpha}{2}}\sigma_{\hat{p}}$
  - Also, valid asymptotically for any estimator resulting as a linear function of a sum of random variables due to CLT

Poisson confidence interval

Another commonly occurring case is where the outcome of a measurement is a Poisson random variable, $f(N, \nu) = \frac{\nu^n e^{-\nu}}{n!}$. In this case:

- $\alpha = P(\hat{\nu} \geq \hat{\nu}_{obs}; a) = 1 - G(\hat{\nu}_{obs}; a) = 1 - \sum_{n=0}^{n_{obs}-1} \frac{a^n e^{-a}}{n!}$
- $\beta = P(\hat{\nu} \leq \hat{\nu}_{obs}; b) = G(\hat{\nu}_{obs}; b,) = \sum_{n=0}^{n_{obs}} \frac{b^n e^{-b}}{n!}$

For one experiment only, we have that $\hat{\nu} = n_{obs}$ is the MLE. Replacing this result in the formulas we can:

- Solve numerically for a, b
- Exploit the Poisson/χ^2 relationship:
  - $a = \frac{1}{2}F^{-1}_{\chi^2_{2n_{obs}}}(\alpha)$ → cdf of a χ^2 distribution with $2n_{obs}$ degrees of freedom computed in α
  - $b = \frac{1}{2}F^{-1}_{\chi^2_{2(n_{obs}+1)}}(1 - \beta)$ → cdf of a χ^2 distribution with $2(n_{obs} + 1)$ degrees of freedom computed in $1 - \beta$

- **Notes:**
  - Important special case is when we want to set an upper limit and we observe $n_{obs} = 0$

$$\beta = \sum_{n=0}^{0} \frac{b^n e^{-b}}{n!} = e^{-b} \rightarrow b = -\log(\beta)$$

Limits near physical boundaries

How to handle estimates near physical boundaries of parameter values?

- This situation is common as experiments often look for new effects, which would imply a given parameter different from zero, e.g. neutrino mass
  - Provide the new parameter's estimate plus a C.I. if data estimate significantly different from zero
  - Report upper limit otherwise (similarly for lower limits)
- What if data estimates lie outside physics allowed regions? E.g. negative mass
  - May happen when estimator is built as a difference due to measurement error, e.g. $\widehat{m^2} = E^2 - p^2$
- Example: if $\hat{\theta} = X - Y$ and $X, Y \sim N$, then $\hat{\theta} \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$ → $\theta_{up} = \hat{\theta}_{obs} + \sigma_{\hat{\theta}} \Phi^{-1}(1-\beta)$
  - This in general corresponds to an interval: $(-\infty, \theta_{up}]$ that surely contains negative values
  - Also, the $\boldsymbol{\theta_{up}}$ **itself may end up being negative!**
  - Statistics interpretation: this is one of the times the interval does not work, but coverage is ensured in the long run
  - **Physicist:** we already knew $\theta > 0$, cannot use negative upper limit as result of expensive experiments!
- Possible options
  - Increase confidence level until the limit enters the allowed region → simply wrong!
  - Shift negative estimates to 0: $\theta_{up} = \max(\hat{\theta}_{obs}, 0) + \sigma_{\hat{\theta}} \Phi^{-1}(1-\beta)$ → often used, but does not maintain coverage
  - Alternatively, Bayesian posterior

More details: Statistical Data Analysis, chapter 9.8

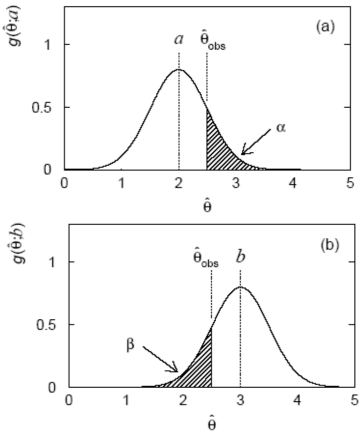
Confidence intervals by inverting a test

A confidence interval can be seen as a hypothesis test:

- Test $H_0: \theta = a$ versus $H_0: \theta < a$ using $\hat{\theta}$ as test statistic
  - Rejection regions is at $\hat{\theta} \geq \hat{\theta}_{obs}$
  - Resulting p-value is α
- C.I.: confidence level α is specified first, and a is a random variable
- Test: null hupothesis H_0: $\theta = a$ is specified first, and $\alpha = p - value$ is a random variable

- Similarly, test $H_0: \theta = b$ versus $H_0: \theta > b$ using $\hat{\theta}$ as test statistic
  - Rejection regions is at $\hat{\theta} \leq \hat{\theta}_{obs}$
  - Resulting p-value is β

- **Note:** the confidence belt can be seen as the acceptance region of the corresponding test



Other methods for building confidence intervals

A confidence interval can be seen as a hypothesis test:

- **Likelihood/χ^2 Method**
  - Use regions where log L decreases by $N^2/2$ from maximum (or χ^2 increases by N^2)
  - For large samples, approximate classical confidence intervals
  - Computationally simpler than exact method
  - Works even for non-Gaussian distributions
- **Profile Likelihood & Likelihood Ratio**
  - Handle nuisance parameters by "profiling"
  - Ratio $\lambda(\theta) = L(\theta,\hat{\nu}\theta)/L(\theta,\hat{\nu})$
  - $-2\log\lambda$ follows χ^2 distribution (Wilks' theorem)
  - Used for hypothesis testing and CI construction
- **Multidimensional Confidence Regions**
- **Feldman-Cousins Unified Approach**
  - Combines hypothesis tests and confidence intervals
  - Uses likelihood ratio ordering principle
  - Handles physical boundaries correctly
  - Avoids "flip-flopping" between upper limits and intervals

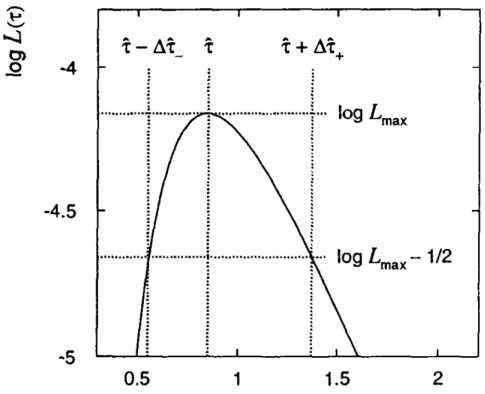


Fig. 9.6 The log-likelihood function $\log L(\tau)$ as a function of τ for a sample of $n = 5$ measurements. The interval $[\hat{\tau} - \Delta\hat{\tau}_-, \hat{\tau} + \Delta\hat{\tau}_+]$ determined by $\log L(\tau) = \log L_{\max} - 1/2$ can be used to approximate the 68.3% central confidence interval.

More details:
Statistical Data Analysis, chapter 9
Backup slides

the image is Likelihood/χ^2 Method.

Bayesian inference

Outline

- Why we need another approach?
- Overview of Bayesian approach
- Examples

Why we need another approach?

Frequentist limitations

Although frequentist approach is solid and widespread, it has several limitations:

- p-value misinterpretation: p-value $\neq P(H)$
- Same for confidence intervals: $P(\theta \in C.I.) \neq 1 - \alpha$
- Struggle to incorporate prior knowledge
- Many theoretic results hold only asymptotically
  - Difficult to deal with small samples
  - Huge statistics needed for rare events
- Model comparison is challenging
  - E.g. LRT does not hold for non-nested hypothesis

Bayesian philosophy

In general, Bayesian statistics is based on a completely inverted conception of randomness:

- **Data is fixed, parameters are uncertain**
- Also, probability is a measure of belief
  - Not «long-run frequency of occurrence» as in frequentist settings
- The whole mechanism is based on:
  - Prior knowledge
  - Bayes theorem as a tool to provide prior updates → prior knowledge is inherently incorporated in our analysis
- Inference is based on the result of the update process: posterior distribution

Bayesian advantages

Given the previous formulation, the Bayesian approach entails several advantages

- Direct probability statements about parameters
- Natural incorporation of prior knowledge

- Better handling of uncertainty and small samples
- Intuitive framework for model comparison
- Alignment with scientific process of updating belief

Bayes' theorem

Bayes' theorem provides a nice mechanism to update probability in light of new evidence:

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

- **Prior Probability, P(A)**: Initial belief before seeing evidence
- **Marginal Likelihood, P(B)**: Overall probability of the evidence
- **Likelihood, $P(B|A)$**: Probability of evidence given the hypothesis
- **Posterior Probability, P(A|B)**: Updated probability after observing evidence

where

where: $P(B) = \Sigma_i P(B|E_i)P(E_i)$

Law of total probability

Key Insights:
- Bayes' Theorem updates beliefs based on new evidence → resembles how we think
- It accounts for both the strength of the evidence and prior knowledge
- In inference, we look at these blocks as distributions!

Prior distribution

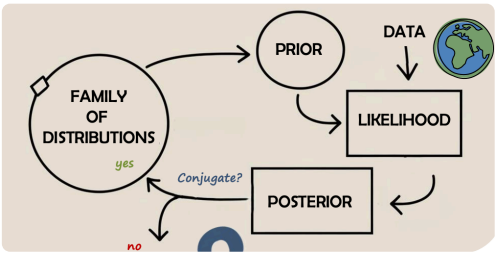
A prior distribution represents our beliefs about the parameters before looking at the data

- Encodes what we know a priori about the possible values of the parameter
- Prior choice is a critical step and it affects our results – Especially crucial when we have limited data!
- Types of prior distributions:
  - Informative
  - Weakly informative
  - Non-informative
  - Conjugate
- Hierarchical priors for hyperparameters
- Important to conduct sensitivity analysis and check robustness to prior choice

Conjugate priors

A prior distribution (ex: probability of disease) is conjugate to a likelihood function if the resulting posterior distribution is in the same probability distribution family as the prior.

- With family we mean the same distribution but different parameters (Poisson(2), Poisson(3))



- Let θ be the parameter of interest
- Prior: $p(\theta)$
- Likelihood: $p(x|\theta)$
- Posterior: $p(\theta|x) \propto p(x|\theta)p(\theta)$
- If $p(\theta)$ and $p(\theta|x)$ are in the same distribution family, $p(\theta)$ is conjugate to $p(x|\theta)$

Key properties:

- Analytical tractability
- Interpretability as prior data
- Sequential updating

Here an example that explains better

Conjugate priors: Beta prior for Binomial data

Let $X \sim Bin(n, p)$ describe the process we are studying. We are interested in the parameter p

- Recall that $f(X = k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$
- What is the conjugate prior for p?
- If we set $p \sim Beta(\alpha, \beta)$, then $f(P = p|\alpha, \beta) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$, where $B(\alpha, \beta)$ is the Beta function
  - To check if $p \sim Beta(\alpha, \beta)$ is conjugate for $X \sim Bin$ we need to derive the posterior
  - By Bayes THM:

$$f(p|x,n) \propto f(X|n,p)f(p) \propto \binom{n}{k} p^k (1-p)^{n-k} \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$
$$\propto p^k p^{\alpha-1}(1-p)^{n-k}(1-p)^{\beta-1} * C \propto p^{k+\alpha-1}(1-p)^{n-k+\beta-1}$$

  - Hence: $f(P = p|x) \sim Beta(k + \alpha, n - k + \beta)$
  - **Note:** in this case it is possible to show that the constant factors nicely combine into $\frac{1}{B(k+\alpha, n-k+\beta)}$, thus returning the exact version of a Beta distribution; however, it is sufficient to derive posterior up to normalization constant

Interpretation:

- The prior Beta(α, β) can be interpreted as $\alpha - 1$ prior successes and $\beta - 1$ prior failures
- The posterior Beta(α', β') incorporates k observed successes and $n - k$ observed failures
- **Update rule**: posterior parameters are simply the prior parameters plus the observed data

Common conjugate prior pairs

We have 6 families of likelihood/conjugate prior pairs:

Prior	Likelihood	Posterior
Beta: $p \sim Beta(\alpha, \beta)$	Bernoulli/Binomial: $X\|p \sim Bin(n, p)$	$p\|X \sim Beta(\alpha + n_S, \beta + n_F)$
Gamma: $\lambda \sim \Gamma(\alpha, \beta)$	Poisson: $X\|\lambda \sim Poisson(\lambda)$	$\lambda\|X \sim \Gamma(\alpha + n_{events}, \beta + n_{obs})$
Normal (mean): $\mu\|\tau \sim N(\mu_0, \tau^2)$	Normal, **known** variance: $X\|\mu \sim N(\mu, \sigma^2)$	$\mu\|X \sim N(\mu', \sigma'^2)$ where: $\mu' = \frac{\frac{\mu_0}{\tau^2} + \frac{\sum X_i}{\sigma^2}}{\sigma'^2}, \quad \sigma'^2 = \frac{1}{\tau^2} + \frac{n}{\sigma^2}$
Inverse-Gamma: $\sigma^2 \sim IG(\alpha, \beta)$	Normal, **unknown** variance: $X\|\sigma^2 \sim N(\mu, \sigma^2)$	$\sigma^2\|X \sim IG(\alpha', \beta')$ where: $\alpha' = \alpha + \frac{n}{2}, \quad \beta' = \beta + \frac{1}{2}\sum(X_i - \mu)^2$
Gamma: $\lambda \sim \Gamma(\alpha, \beta)$	Exponential: $X\|\lambda \sim Exponential(\lambda)$	$\lambda\|X \sim \Gamma(\alpha + n, \beta + \sum X_i)$
Dirichlet: $\boldsymbol{\theta} \sim Dirichlet(\alpha_1, \dots, \alpha_K)$	Multinomial: $X\|\boldsymbol{\theta} \sim Multinomial(n, \boldsymbol{\theta})$	$\boldsymbol{\theta}\|X \sim Dirichlet(\alpha_1 + X_1, \dots, \alpha_K + X_K)$

Jeffrey's prior

Jeffrey's priors are a way of expressing «objective» or «non-informative» prior knowledge, i.e. let the data speak for themselves.

This mean that the prior should have little or no prior information about the parameter of interest.

- **Proportional to Fisher information matrix:** $\pi(\theta) \propto \sqrt{\det(I(\theta))}$, where $I(\theta) = E[(-\partial^2/\partial\theta^2) \log p(x|\theta)]$
- **Note:** invariant under reparameterizations (e.g. $\tau = 1/\theta$)
  - This is not guaranteed in general, e.g.: for $\theta \sim Unif(0,1)$ then $\tau = \theta^{-1} \nsim Unif(0,1)$ (see water and wine paradox)
- Some common cases:
  - Poissonian mean: $p(\mu) \propto 1/\sqrt{\mu}$
  - Poissonian mean with background b: $p(\mu) \propto 1/\sqrt{\mu + b}$
  - Gaussian mean: $p(\mu) \propto 1$
  - Gaussian r.m.s: $p(\sigma) \propto 1/\sigma$
  - Binomial parameter: $p(\varepsilon) \propto 1/\sqrt{\varepsilon(1-\varepsilon)}$
- **Jeffrey's priors are often improper** (do not integrate to one)
  - Not a problem as long as the posterior does!

Bayesian point estimation

How do we summarize the posterior distribution? We have many options!

- **Posterior mean:** $E[\theta|x] = \int \theta p(\theta|x) d\theta$ → integrate wrt θ!
  - Minimizes squared error loss
  - Often easy to compute analytically for conjugate priors
- **Posterior Median:** use θ_{med} such that $P(\theta \leq \theta_{med}) = 0.5$
  - Minimizes absolute error loss
  - Robust to outliers
  - Invariant under monotonic transformations
  - Often requires numerical computation
- **Posterior Mode (MAP Estimate):** $\theta_{MAP} = \text{argmax}_\theta\{p(\theta|x\}$
  - Maximizes the posterior density
  - Often similar to MLE for large samples and flat priors
  - Not invariant under reparameterization
  - Can be computed via optimization

- **Note:** importantly, we now have the whole distribution so we can compute whatever quantity of interest (e.g. 25th, 75th percentiles, variance, skewness, …)

Choosing a Point Estimate

There are several factors that influence how we choose Bayesian point estimates:

- Depends on the use-case
  - Ease of computation
  - Interpretability
  - Robustness
  - Invariance properties
- Comparison to Frequentist Estimates
  - MLE often similar to MAP with flat prior
  - Bayesian estimates incorporate prior information
  - Bayesian framework provides natural uncertainty quantification
- Limitations:
  - Can be misleading for multimodal posteriors

Credible intervals

A credible interval is a range of values that contains the true parameter value with given posterior probability

- **Interpretation**: "There is a α% probability that the true parameter value lies within this interval, given the data and our prior beliefs"

Types of Credible Intervals

- Central credible interval of size α
  - An interval a, b where $P(\theta < a|data) = P(\theta > b|data) = 1 - \alpha/2$
  - Easy to compute and interpret
  - May not be the shortest possible interval
- Highest Posterior Density (HPD) Interval •
  - The shortest interval containing α% of the posterior probability
  - Always includes the posterior mode
  - May be disjoint for multimodal posteriors
  - Invariant under one-to-one transformations of parameters

Computed analytically (if we know $f(\theta|x)$) or numerically.

Limitations:
- Depend on prior specification
- Actual coverage may be greater/lower than nominal value

Hypothesis testing

The Bayesian approach gives a direct way of measuring the probability of hypothesis

- This time we can compute hypotheses probability directly: $P(H_0|data)$ and $P(H_1|data)$
- Then the test is based on the Bayes factor:

$$BF = \frac{P(data|H1)}{P(data|H0)} = \frac{P(H1|data)P(H1)}{P(H0|data)P(H0)}$$

- BF > 1 means that data more strongly support $H1$
- BF < 1 means that data are more compatible with $H0$
    - Bayes factors quantify evidence in favor of a null hypothesis, rather than only allowing $H0$ to be rejected or not

Sampling approaches

However, closed analytical solutions are not always available → resort to sampling

- **Metropolis-Hastings Algorithm**

– Basic principle: Proposal and acceptance/rejection
– Key steps:
• Propose a new state
• Calculate acceptance probability
• Accept or reject the proposal
– Tuning the proposal distribution

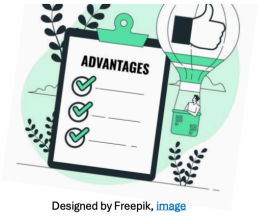
- **Gibbs Sampling**

– Sampling each parameter conditionally on others
– Useful for hierarchical models
– Convergence properties
• Other methods:
– Importance sampling
– Variational inference

Advantages: Bayesian VS Frequentist

Bayesian

- Intuitive interpretation of results
- Natural incorporation of prior information
- Handles small sample sizes better
- Straightforward approach to complex models
- No need for p-values or adjustments for multiple comparisons
- Provides full posterior distribution



Designed by Freepik, image

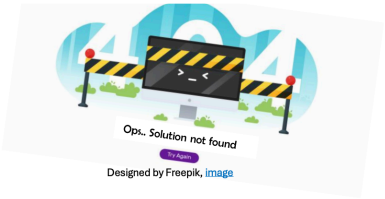
Frequentist

- Objectivity (no prior specification needed)
- Well-established procedures and software
- Often computationally simpler
- Long-run performance guarantees
- Widely accepted in many scientific fields

Challenges: Bayesian VS Frequentist

Bayesian

- Prior specification can be subjective
- Can be computationally intensive
- Can be sensitive to prior choice with small samples



Frequentist

- Interpretation of p-values and significance
- No direct probability statements about hypotheses
- Complex models and small sample sizes
- Multiple comparisons and p-hacking

Practical considerations

The choice may often depend on practical requirements and considerations:

• Field-specific conventions

• Nature of the problem, e.g.:

– Availability of prior information

– Possibility to run repeated experiments

• However, often both approaches reach similar results

– Large sample sizes

– Objective priors can lead to similar frequentist methods – Calibrated Bayes approaches attempt to ensure good frequentist properties

Learning theory

What is learning?

Learning is about creating systems that can improve their performance on a task through experience (i.e., data). More formally:

• **Experience (E):** Data used for learning
– Historical observations
– Experimental measurements
– Simulated data

• **Task (T):** What we want to accomplish
– Predicting house prices
– Classifying particle interactions
– Clustering galaxy types

• **Performance (P):** How we measure success
– Prediction accuracy
– Mean squared error
– Classification precision/recall

Learning paradigms & supervision

Goal:

Learning a mapping function , where:
• X: Input space (e.g., raw data, signals, measurements).
• Y: Output space (e.g., predictions, classifications, clusters).

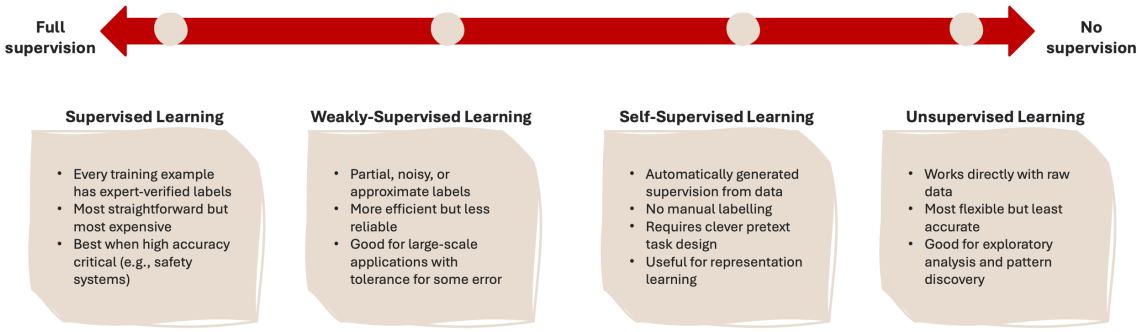
Labels & Annotations:
Annotations (e.g., labels for target outputs) guide the learning process but come with challenges:

- Require expertise.
- Are time-intensive.
- Demand a large volume for effectiveness.

Examples: Associating particle species with signals, associating invariant mass with measured momenta.

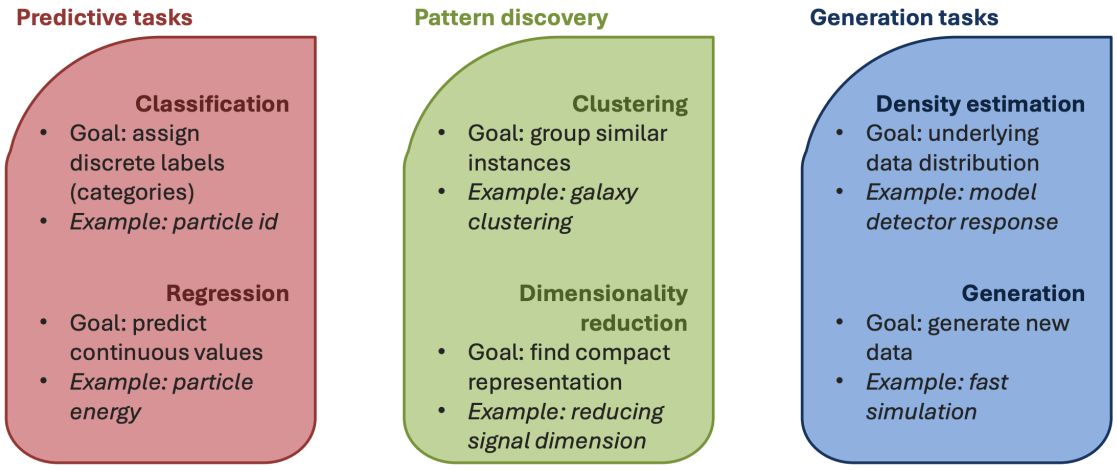
Supervision Spectrum

We have several learning paradigms depending on the degree of supervision provided to the model



Learning tasks: what are we trying to learn?

We have different learning tasks depending on what kind of output the model should produce

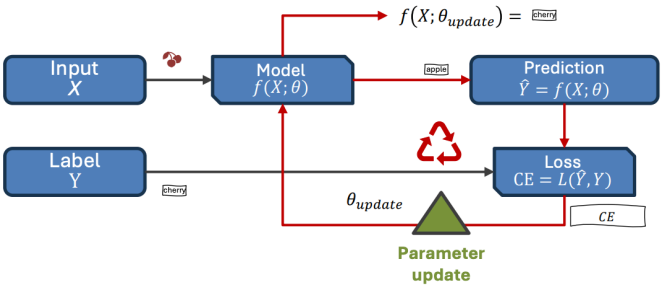


Training procedure

The learning phase, training, is carried out differently depending on the learning paradigm

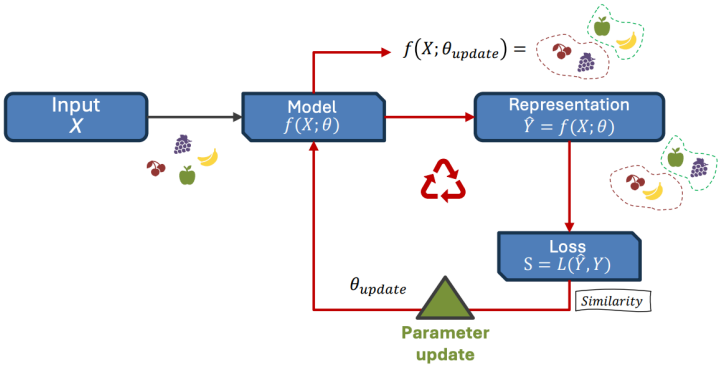
Supervised

• The model learns from pairs of inputs/desired outputs;
– e.g. fruit image/fruits category
• Parameters adjusted to minimize "difference" between predicted and desired outputs (loss function)
– E.g. Cross Entropy (CE)
• Learning is guided by expert annotation
• Performance clearly defined by the loss or other quantitative measures
– E.g. accuracy, precision, recall



Unsupervised

- The model sees data only and searches patterns in data
  - E.g. fruit image/groups of similar images
- Parameters adjusted to compute convenient representation
  - E.g. latent space where similar images are close-by
- Learning is guided by data structure and representation, without feedback
- Not clear how to measure performance
  - Typically requires interpretation of results
  - Many interpretations may be possible!!
- which one is of interest?



Remarks

Loss Function Selection

- Defines what the model considers as "error"
- **Must align with physical objectives**
  - Classification: Cross-entropy for probabilities
  - Regression: MSE for continuous values
  - Custom losses for physics constraints
- Influenced by:
  - Nature of the data (discrete/continuous)
  - Noise characteristics and challenges
  - Domain-specific requirements

Supervised Learning Challenges

- **Overfitting**: memorizing vs. learning
  - performs well on training data but fails to generalize
  - **Data splitting** strategy
    - Training: learn parameters
    - Validation: tune hyperparameters
    - Test: final evaluation
    - Must be **independent samples**!

Unsupervised Learning

- Domain knowledge incorporation
  - **Definition of similarity**
  - Choice of data representation
  - Physical constraints
  - Selection of relevant features
- Links to dimensionality reduction
  - Balance between compression and information

Practical Tips

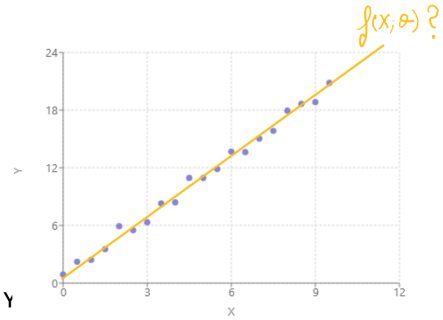
- **Start simple, add complexity** as needed
- **Validate against physical intuition**
- Consider computational resources
- **Document** assumptions and choices
  - Experiment tracking and MLOps

Least Squares

Linear regression via Least Squares

Least Squares (LS) are an estimation technique adopted to solve a linear regression task:

- Regression is a learning task aiming at predicting continuous values
  - X is the independent (control) variable (input, feature, observable)
  - Y is the dependent variable (output, target)
  - We want to find a function f: X → Y such that $Y = f(X;\theta)$
  - If $f(X;\theta)$ is a **liner function of the parameters** θ, then we have linear regression



- Problem setup
  - Imagine we observe a sample of (X,Y) pairs: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
  - Goal: find $f(X;\theta)$ that best approximates the relationship between X and Y
  - Linear model: $f(X;\boldsymbol{\theta}) = \beta_0 + \beta_1 X$, where $\boldsymbol{\theta} = (\beta_0, \beta_1)$
  - How do we determine the optimal value for the vector $\boldsymbol{\theta}$**?**

Least Squares method resolves this problem by minimizing the Mean Square Error (MSE)

- Define the error (loss) in terms of MSE between predicted and true values of Y:

$$MSE = \sum_i (y_i - \hat{y}_i)^2$$

- Why MSE?
  - Heavier penalization to larger differences
  - Positive and negative errors are treated equally
  - Mathematically convenient (differentiable)
- Objective: minimize total observed MSE wrt β_0, β_1
  - $MSE = Loss(\beta_0, \beta_1) = \sum_i \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2$
  - Taking partial derivatives
    - $\frac{dLoss}{d\beta_0} = -2\ \sum_i \big(y_i - (\beta_0 + \beta_1 x_i)\big) \rightarrow \widehat{\beta_0}^{LS} = \bar{y} - \beta_1 \bar{x}$, where we replace $\beta_1 = \widehat{\beta_1}^{\mathrm{LS}}$
    - $\frac{dLoss}{d\beta_1} = -2\ \sum_i \big(y_i - (\beta_0 + \beta_1 x_i)\big)\, x_i \rightarrow \widehat{\beta_1}^{\mathrm{LS}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{COV(x,y)}{VAR(x)}$
- $\widehat{\beta_0}^{\mathrm{LS}}, \widehat{\beta_1}^{\mathrm{LS}}$ are said LS estimates of the coefficients for linear regression

Least Squares estimates have some nice properties:

- LS provides **Best Linear Unbiased Estimator** (BLUE) for $\boldsymbol{\theta}$ (Gauss-Markov theorem)
  - Unbiased and least sampling variance
- If errors are normally distributed, then **LS are equivalent to MLE**
  - We can write: $Y = f(X;\boldsymbol{\theta}) + \epsilon = \beta_0 + \beta_1 X + \epsilon$, where $\epsilon = Y - \hat{Y}$ is the error (residuals)
  - If $\epsilon \sim N(0, \sigma^2)$, then also $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$
  - It follows that the log-likelihood can be written as:

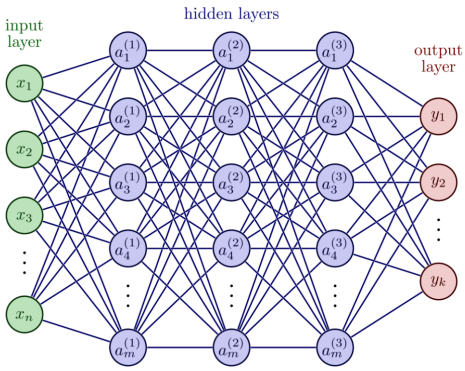
$$\ell(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_i (Y_i - \beta_0 - \beta_1 X_i)^2$$

  - Note that $\ell(\beta_0, \beta_1, \sigma^2) \propto MSE$ when we are interested just in β_0, β_1
    → Least Squares are equivalent to Maximum Likelihood estimates for β_0, β_1

Neural Networks and Decision Trees

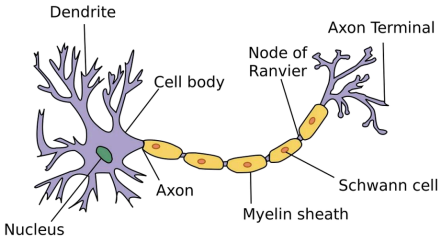
Neural Networks (NN) are a class of models that have proven effective at learning from data

- Powerful universal function approximators
- Inspired by biological neural systems
- Very effective for:
  - Pattern recognition
  - High-dimensional data
  - Non-linear relationships
- Widely applied in physics as well:
  - Particle identification
  - Event reconstruction
  - Fast simulation

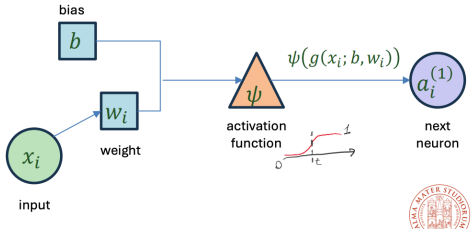


The basic unit of a Neural Network is the "**artificial neuron**", also called "**perceptron**"

- Inspired by biological neurons
  - The cell body receives a signal and process it
  - If the signal is interesting, the cell body gets excited
  - When the excitement exceed a given threshold, the axons activate and transmit the signal to neighboring cells

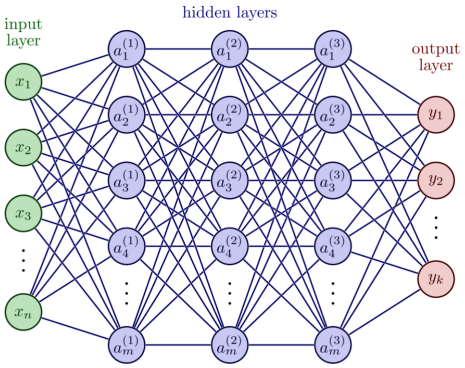


- Artificial neurons are a mathematical representation of the above behavior:
  - The signal is represented by the **input**, x_i
  - The excitement level is formulated as a **linear transformation of the signal:** $g(x_i; b, w_i)$
  - The result is **passed through an activation function,** ψ, that mimics the activation mechanism
    - If the processed signal overcomes the activation threshold, it is passed through the next neurons

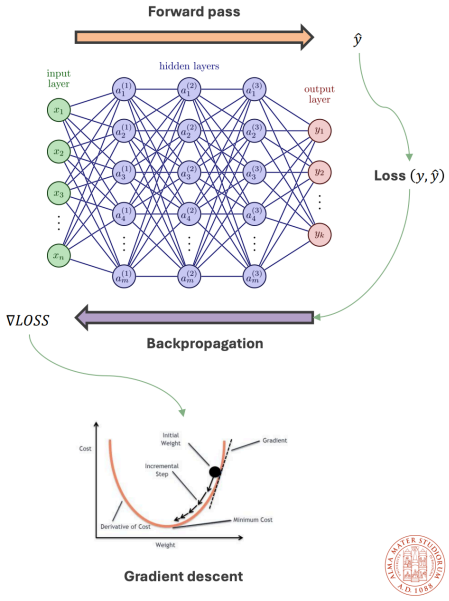


Building upon artificial neurons, we can design custom network architectures formed by the following blocks:

- **Input layer**
  - Collect all raw data inputs
- **Hidden layers**
  - Possibly more than one
  - The more we add, the deeper the architecture
  - Each neuron is connected to all neurons in adjacent layers, i.e. Fully Connected Neural Network (FCNN)
- **Output layer**
  - Final result, network prediction
  - The number of neurons depend on the task
    - Classification: neurons = n. of classes
    - Regression: typically one neuron



Learning phase



Gradient descent

The learning phase can be break down into 4 steps:

- Forward pass:
  - Input is propagated to the network to get a prediction
  - Weights and biases are initialized "somehow"
- Loss computation
  - The loss function is evaluated to measure prediction error
- Backpropagation
  - Compute gradients of loss wrt parameters
  - Chain rule allows to retrieve these gradients for all layers proceeding backwards: from output to previous layer an so on…
- Parameter update
  - Gradient descent optimization
  - Learning rate, α, controls the update "size"
  - Update rule: $w_{new} = w_{old} - \alpha \nabla LOSS$

Universal Approximation Theorem

The Universal Approximation Theorem (UAT) stats that

> "A feedforward network with a single hidden layer containing a finite number of neurons can approximate any continuous function on compact subsets of $\mathbb{R}^n$, under mild assumptions about the activation function"
>
> K. Hornik, M. Stinchcombe, H. White, *Multilayer feedforward networks are universal approximators*(1989)

Implications

- A NN with a single hidden layer (Single Layer Perceptron, SLP) can approximate any function, provided that:
  - Sufficient number of hidden neurons
  - Appropriate activation function (introduce non-linearity)
  - Weights and biases are learned correctly
- However, this theorem just shows the existence!
  - Do not specify how to find correct weights
  - Number of hidden neurons might be impractically large (computationally unfeasible)

If a SLP can learn any function, then why bothering with deeper architectures?

Deep architectures have several advantages:

- Hierarchical Feature Learning
  - Lower layers: basic features
  - Middle layers: feature combinations
  - Upper layers: abstract representation
- More efficient learning than shallow architectures
  - Fewer total parameters needed for same expressivity
  - Better generalization properties
  - More efficient training and inference
- Empirical Success
  - Consistently better performance in practice
  - More robust feature learning
  - Better transfer learning capabilities

Key Takeaway

While UAT shows that shallow networks are theoretically sufficient, deep networks are practically superior due to:

- More efficient parameter usage
- Better representation learning
- Natural hierarchy matching real-world processes

Training remarks

Challenges:

- **Overfitting** is very common if architecture not tuned properly
- Mitigation strategies: Regularization, Dropout, Early stopping
- **Vanishing/exploding gradients**
- Activation functions squeeze neuron outputs between 0 and 1

If architecture is deep, we risk ending up with repeated multiplications of very small numbers, which lead either to vanishing gradients (similar for exploding case when activation is unbounded, e.g. relu)

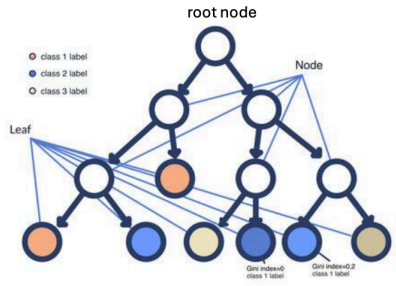
Mitigation strategies: careful weight initialization, batch normalization, residual connections

- **Practical aspects to mind**
- Batch size selection
- Learning rate initialization and scheduling
- Loss function

Decision Trees

Decision trees are based on a 3 structural components and a splitting criterion:

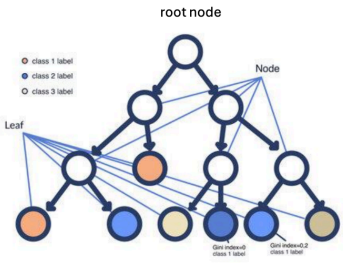
- Root node:
  - Top of the tree, it contains all of the data together
  - The sequential splitting starts here by selecting the most discriminative feature
- Internal nodes
  - Decision points for the growth of the tree
  - Contain subsets of data determined by conditions imposed by previous splits
- Leaf nodes
  - Terminal nodes of the tree
  - Contain homogeneous subsets of data
  - Predictions are retrieved by applying a function to each leaf separately
    - Classification: majority class
    - Regression: mean/median value



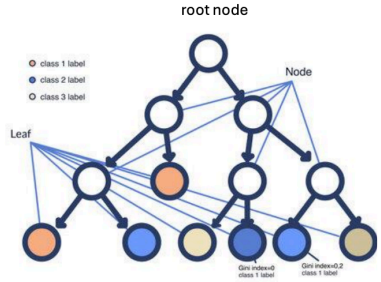
- Splitting criterion
  - Criterion used to select
    - Most discriminative feature
    - The cutoff value for binary split
  - Classification
    - Gini impurity or Entropy
  - Regression
    - MSE, Mean Absolute Error (MAE)

Decision Trees are trained by a recursive algorithm. It starts from the root node and **repeatedly splits data to create more homogeneous sub-nodes**. The procedure is repeated **until a convergence criterion is met**. The final nodes are called leaves, and they should contain homogeneous subsets of data.

- At each iteration, all nodes and all features are scanned
- For a fixed feature and node:
  - Attempt splitting node data by applying a binary cut to that feature, e.g. split observations at Energy > 50 GeV
  - Measure how good the split is, i.e. we quantify entropy/impurity or loss, *L*, derived by the split
- The procedure is repeated for varying cutoffs, for all combinations of features and existing nodes
- At the end of the iteration, all impurity/loss metrics are compared, and the split corresponding to the best metric is applied
- Repeat iteration searching for new splits until a stopping criterion is met
  - Max depth, minimum observations per leaf, minimum gain, …



Imagine we are studying a particle identification problem, where we have:



- Features (X): Energy, Track length and Shower width
  - These are the observables of our sample
- Target variable (Y): Muon, Pion, Electron
  - These represent the particle species we are studying

- The training of a Decision Tree consists of repeatedly splitting our data into homogeneous subgroups
  - easier to distinguish particle species

```
Root: Energy > 50 GeV?
├── Yes: Track length > 10 cm?
│   ├── Yes: Muon
│   └── No: Pion
└── No: Shower width > 5 cm?
    ├── Yes: Electron
    └── No: Further splits...
```

Building upon these foundations, researcher have developed several tweaks to improve the learning result:

- Pruning
  - If we let the model grow indefinitely, then it will overfit
  - Hence, we must tune appropriately the convergence criteria to prevent this
  - Pruning is an alternative approach that avoid setting stringent convergence conditions
    - let the tree grow and explore many branches
    - at the end, remove branches that do not improve validation performance
  - Effective in combatting overfitting
  - Improves generalization
  - Reduces model variance
- Popular extensions: Ensembles
  - Random forests: use multiple trees trained on random subsets of the features; the final prediction is retrieved as voting/average of those trees
  - Gradient Boosting: build strong predictor from sequence of weak learners; each new model corrects errors of previous ensemble