Fondamenti di Analisi dei Dati

from data analysis to predictive techniques

Prof. Antonino Furnari (antonino.furnari@unict.it)

Corso di Studi in Informatica

Dip. di Matematica e Informatica

Università di Catania

Università di Catania

Probability for Data Analysis

A framework for reasoning about uncertainty in data science

What values will variables assume?

When doing data science, we often ask "what values will variables assume"? We have two main types of systems:

Deterministic Systems

Perfect predictability from initial conditions—like calculating object speed using Newtonian physics.

Examples: planetary motion, chemical reactions with known conditions, or simple mechanical systems.

Uncertain Systems

Cannot predict outcomes with certainty—require probabilistic reasoning and statistical methods.

Examples: stock market prices, weather forecasting, or medical diagnosis outcomes.

Sources of Uncertainty

In many real-world scenarios, we face **uncertainty**. Modeling the relationship between variables in a deterministic way is not always possible. This uncertainty can arise from several sources:

Inherent Randomness

Truly stochastic events like rolling a die or tossing a coin. Even with perfect knowledge, the outcome is unpredictable.

Incomplete Observability

Missing information, like the Monty Hall problem: the game is deterministic, but the player lacks full information—so the outcome feels uncertain.

Incomplete Modelling

Limited tools to model complex systems fully. For example, **a robot with only an RGB camera** may estimate object positions in 2D, but cannot reconstruct full 3D coordinates. The missing dimensions introduce uncertainty.

Why Probability Theory Matters

Probability theory provides a **consistent framework** for reasoning about uncertainty. It allows us to:

Quantify Events

Assign numerical probabilities to uncertain occurrences, making them measurable.

Combine Information

Integrate both known data and unknown variables to form a complete picture.

Derive Statements

Use formal rules to deduce new, logically sound probabilistic conclusions.

Make Predictions

Forecast future outcomes and assess risks even when information is incomplete.

Random Experiments

In practice, when data acquisition is affected by uncertainty, we refer to the process as a **random experiment**. Informally, we define a random experiment as:

> An experiment that can be repeated any number of times, potentially leading to different outcomes.

We introduce the following terminology:

Sample Space Ω

Set of all possible outcomes

Example, in the random experiment of **rolling a die**, the sample space will be $\Omega = \{1, 2, 3, 4, 5, 6\}$

Simple Event ω_i

A single possible outcome

When rolling a die, a simple event may be: $\omega_1 = 1$ (rolling a 1)

Event $A \subseteq \Omega$

A subset of the sample space

When rolling a die, events may be:

- $A = \{1\}$ (rolling a 1)
- $A = \{2, 4, 6\}$ (rolling a even number)
- $\bar{A} = \{1, 3, 5\}$ (rolling a odd number) - this is also called "the complement" of A

Random Variables

Random variables are similar to statistical variables, but they are used when dealing with uncertainty. We will define a random variable as follows:

> A random variable is a variable whose values depend on outcomes of a random phenomenon

Formally, it is a function

$$X : \Omega \to E$$

where E is a measurable space

Random variables are often denoted by **capital letters** (e.g., X, Y, Z). As with statistical variables, random variables can be of different types.

	Discrete	**Continuous**
Scalar	Tossing a coin	Height of a person
Multidimensional	Pair of dice	Coordinates of a car

Titanic Dataset Example

Random experiment: randomly picking a passenger from the ship

Random Variables

C: Passenger class {1, 2, 3}

S: Sex {male, female}

Questions We'll Answer

Is it more likely to pick a woman or a man?

If I pick from 1st class, what sex is more likely?

Titanic Toy Example

We will consider a toy example derived from the Titanic dataset. The underlying random experiment will be as follows:

1

Randomly pick a class

Choose from first, second, or third class

2

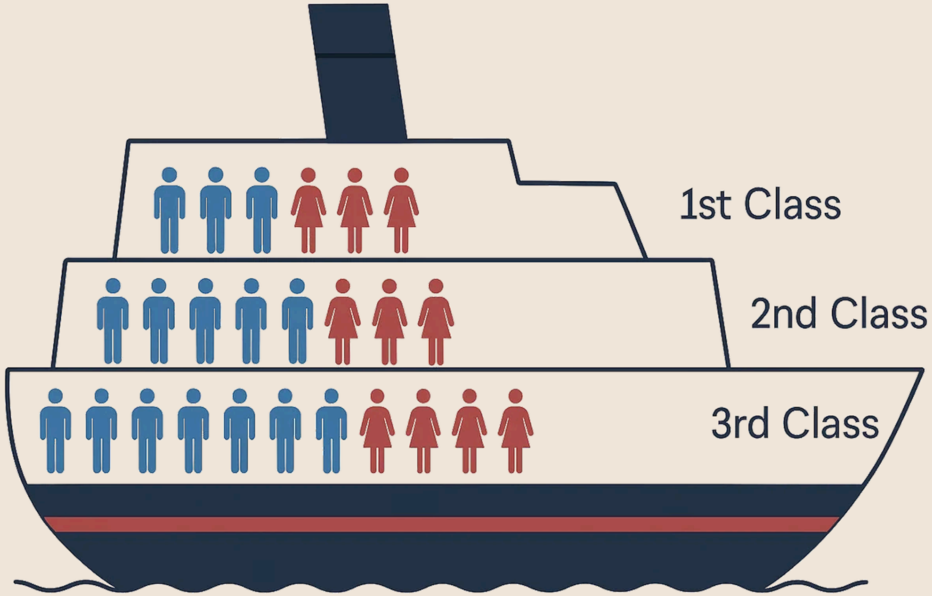
Randomly pick a passenger

Select one passenger from that class

3

Observe and replace

Record the passenger's sex, then return them to the pool



We define two random variables:

- C: passenger class
- S: passenger's sex

For instance, if we pick a passenger, we may obtain: $C = 1$, $S = male$.

What is Data?

We will now consider a working definition of data which is linked to probability theory:

Data = The values assumed by a random variable

- Example: S = male *is data*
- The data is the **pair** <random variable, value>, not just the value
- S = male represents an **event**: "I randomly picked a passenger and their sex was male"
- "male" would not have a definite meaning if not associated with a random variable

Probability Notation

When working with data, certain events are more likely to appear than others. For instance, in our toy example, if I take a random passenger, **which class is more likely to be observed**?

We quantify this concept with the one of **probability**.

A probability is a number between 0 and 1 associated to an event, where

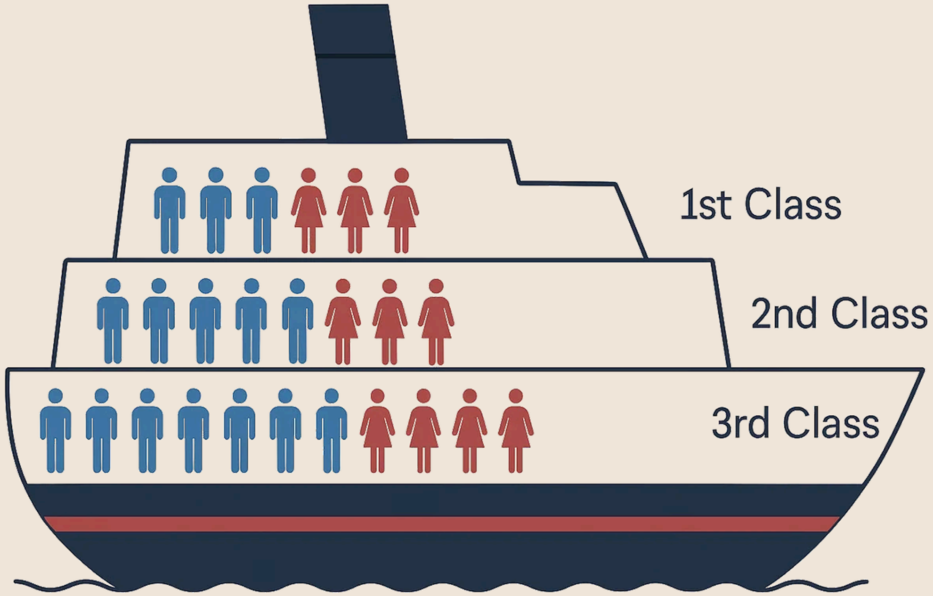
- **0 means impossible to be observed**
- **1 means certain to be observed**

We denote probability with a capital P as follows

$$P(C = 1) = \text{probability of picking from first class}$$

When context is clear, we may omit the variable:

$$P(male) = P(S = male)$$



Kolmogorov's Axioms and Corollaries

Kolmogorov in 1933 provided three axioms which define the "main rules" that probability should follow:

Axiom 1: Non-negativity

$0 \leq P(A) \leq 1, \forall A \subseteq \Omega$

Axiom 2: Certainty

$P(\Omega) = 1$

Axiom 3: Additivity

If $A \cap B = \varnothing$, then $P(A \cup B) = P(A) + P(B)$

These corollaries follow from the axioms:

Complement Rule

$$P(\overline{A}) = 1 - P(A)$$

Impossible Event

$$P(\emptyset) = 0$$

General Addition Rule

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

Monotonicity

$$A \subseteq C \Rightarrow P(A) \leq P(B)$$

Laplace Probability

When all outcomes in a random experiment are considered equally probable, this is called a **Laplace experiment.** In this case, we can calculate the probability of a given event as the ratio between the favorable outcomes and the possible outcomes:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{favourable outcomes}}{\text{possible outcomes}}$$

Die Rolling Example

P(rolling a 3) = 1/6

P(even number) = 3/6 = 1/2

Estimating Probabilities from Observations: Frequentist Approach

According to the frequentist approach, we can estimate probabilities by repeating an experiment for a large number of times and then computing:

$$P(X = x) = \frac{\#\ \text{times}\ X = x}{\#\ \text{trials}}$$

- [ ] If we toss a coin 1,000 times and get 499 heads, then P(head) ≈ 0.499

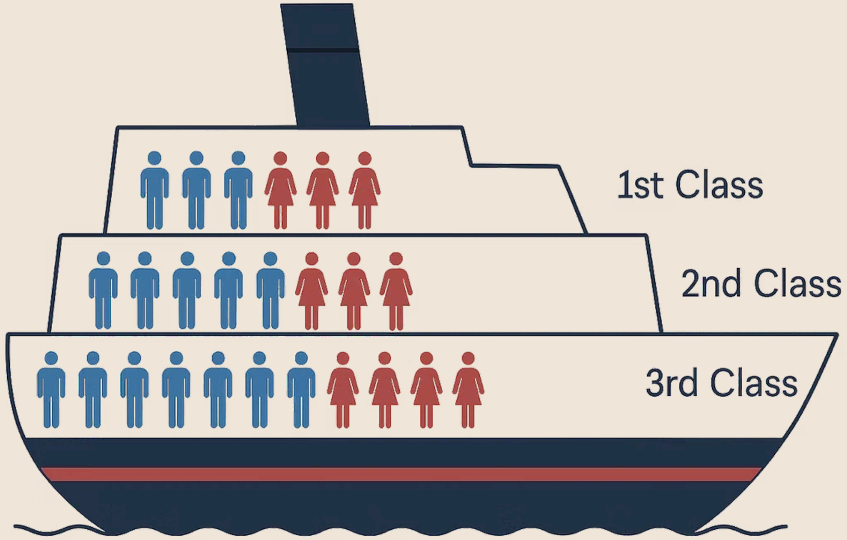
Bayesian Approach to Probability

The Bayesian approach to probability offers a different perspective on probability theory. It sees probability as a **measure of belief or uncertainty**.

Incorporates **prior knowledge** and updates beliefs with new information

Particularly useful for unique events where frequency analysis doesn't apply

- [ ] Example: "What is the probability that the sun will extinguish in 5 billion years?"



Probability Example

Suppose we randomly draw 25 passengers and observe:

- first class: 6 times
- second class: 8 times
- third class: 11 times

With a frequentist approach we can estimate the following probabilities:

- $P(C = 1) = \frac{6}{25} = 0.24$
- $P(C = 2) = \frac{8}{25} = 0.32$
- $P(C = 3) = \frac{11}{25} = 0.44$

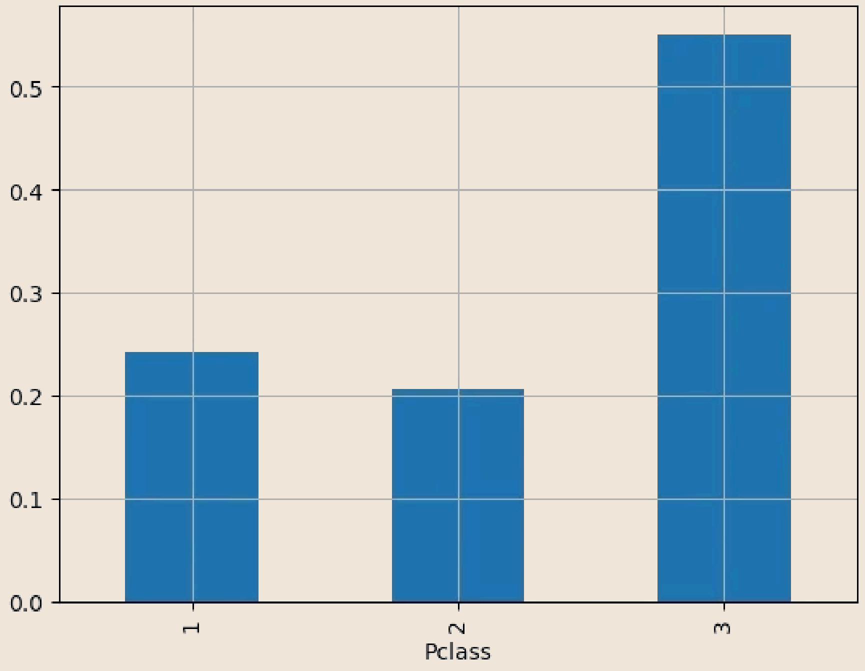
Computing Probabilities from Data

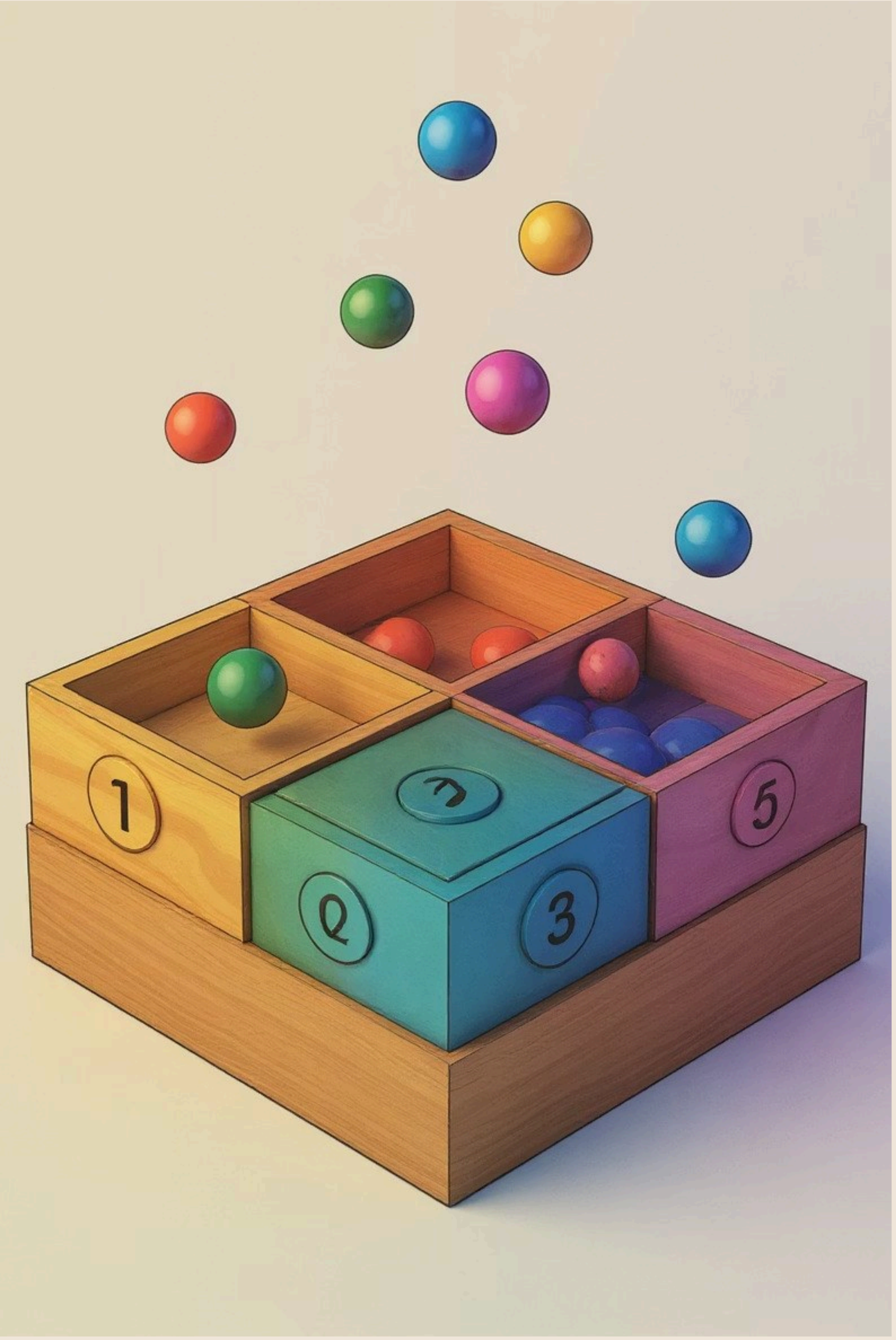
In frequentist terms, probabilities are the same as relative frequencies we have seen in the previous lectures.

For instance:

```
titanic['Pclass'].value_counts(normalize=True).sort_index()
```

```
Pclass
1    0.242424
2    0.206510
3    0.551066
Name: proportion, dtype: float64
```

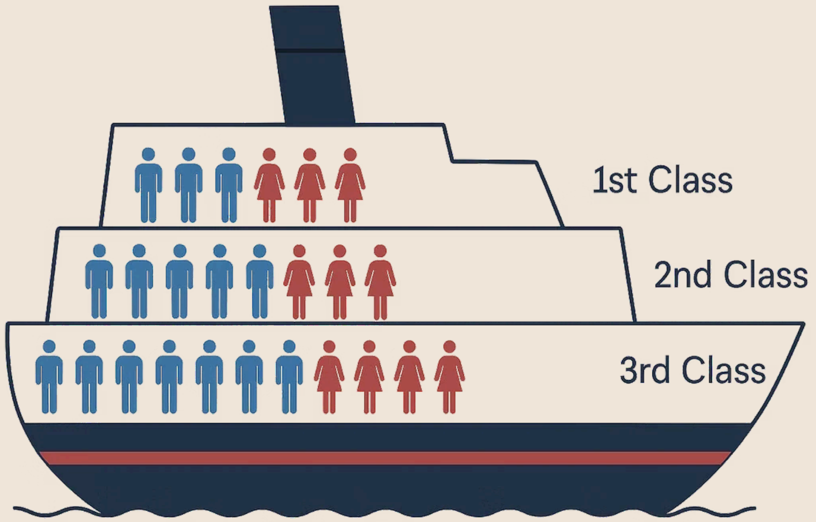




Joint Probability

- Probability of **multiple variables** simultaneously: P(C, S)
- Joint probabilities are **symmetric**: P(X, Y) = P(Y, X)
- We can have joint probabilities of arbitrary number of variables: $P(X_1, X_2, \ldots, X_n)$
- Example: P(C=1, S=male) represents the probability of picking a male passenger from first class
- When dealing with multidimensional variables, we have joint probabilities:

$$X = [X_1, X_2]$$
$$P(X) = P(X_1, X_2)$$



Joint Probability Example

Let's build a contingency table:

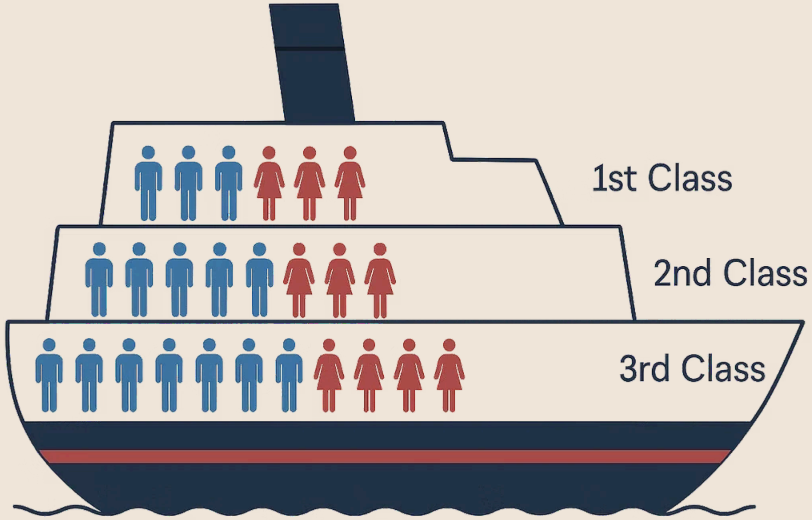
	First Class	Second Class	Third Class	All
male	3	5	7	15
female	3	3	4	10
All	6	8	11	25

We can easily derive a joint probability with the frequentist approach:

$$P(C = 1, S = male) = \frac{\#\text{male occurs in first class}}{\#\text{all passengers}} = \frac{3}{25} = 0.12$$

We can also obtain univariate probabilities using the "All" column:

$$P(C = 1) = \frac{15}{25} = 0.6$$



Joint Probability Table

We can obtain the joint probability table by dividing the contingency table by 25:

	First Class	Second Class	Third Class	All
male	3/25	5/25	7/25	15/25
female	3/25	3/25	4/25	10/25
All	6/25	8/25	11/25	25/25

Alternatively, with decimal values:

	First Class	Second Class	Third Class	All
male	0.12	0.2	0.28	0.6
female	0.12	0.12	0.16	0.8
All	0.24	0.32	0.44	1

Computing a Joint Probability Table

```
pd.crosstab(titanic['Sex'], titanic['Pclass'])
```

Sex \ Pclass	1	2	3
female	94	76	144
male	122	108	347

```
pd.crosstab(titanic['Sex'], titanic['Pclass'], margins=True)
```

Sex \ Pclass	1	2	3	All
female	94	76	144	314
male	122	108	347	577
All	216	184	491	891

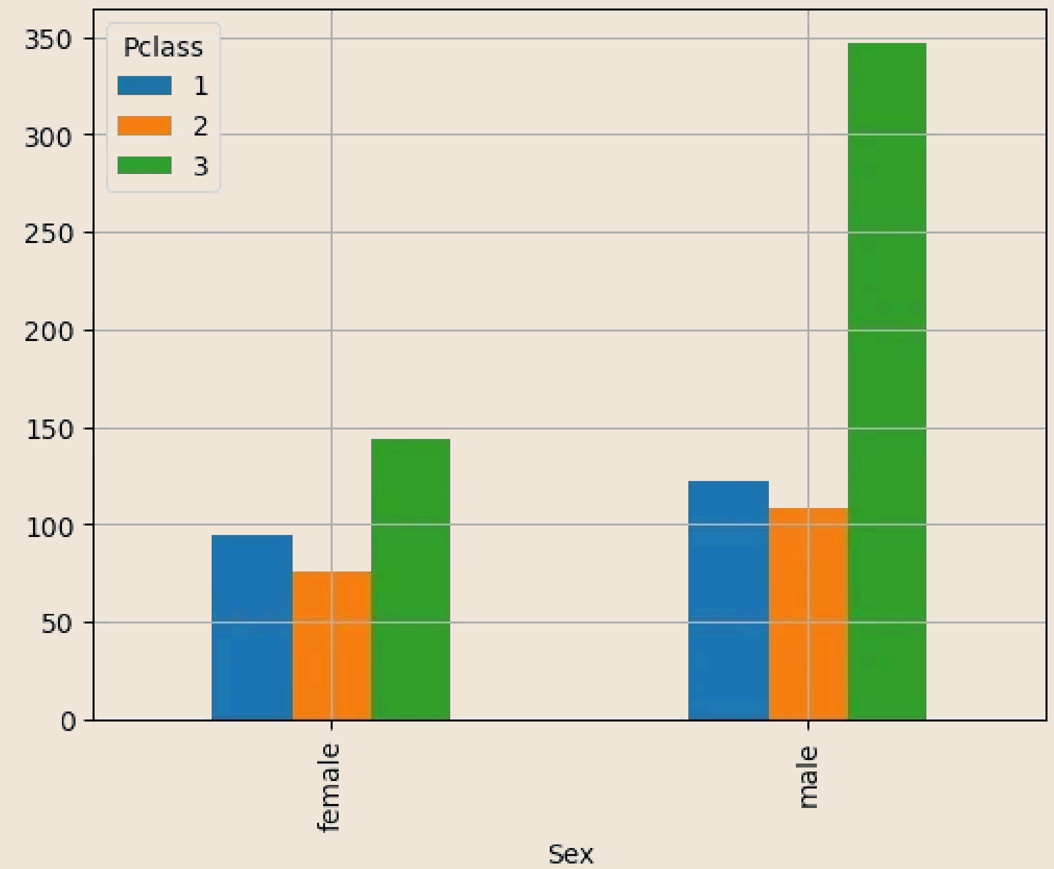
```
pd.crosstab(titanic['Sex'], titanic['Pclass'], normalize=True, margins=True)
```

Sex \ Pclass	1	2	3	All
female	0.105499	0.085297	0.161616	0.352413
male	0.136925	0.121212	0.389450	0.647587
All	0.242424	0.206510	0.551066	1.000000

Graphical Representations

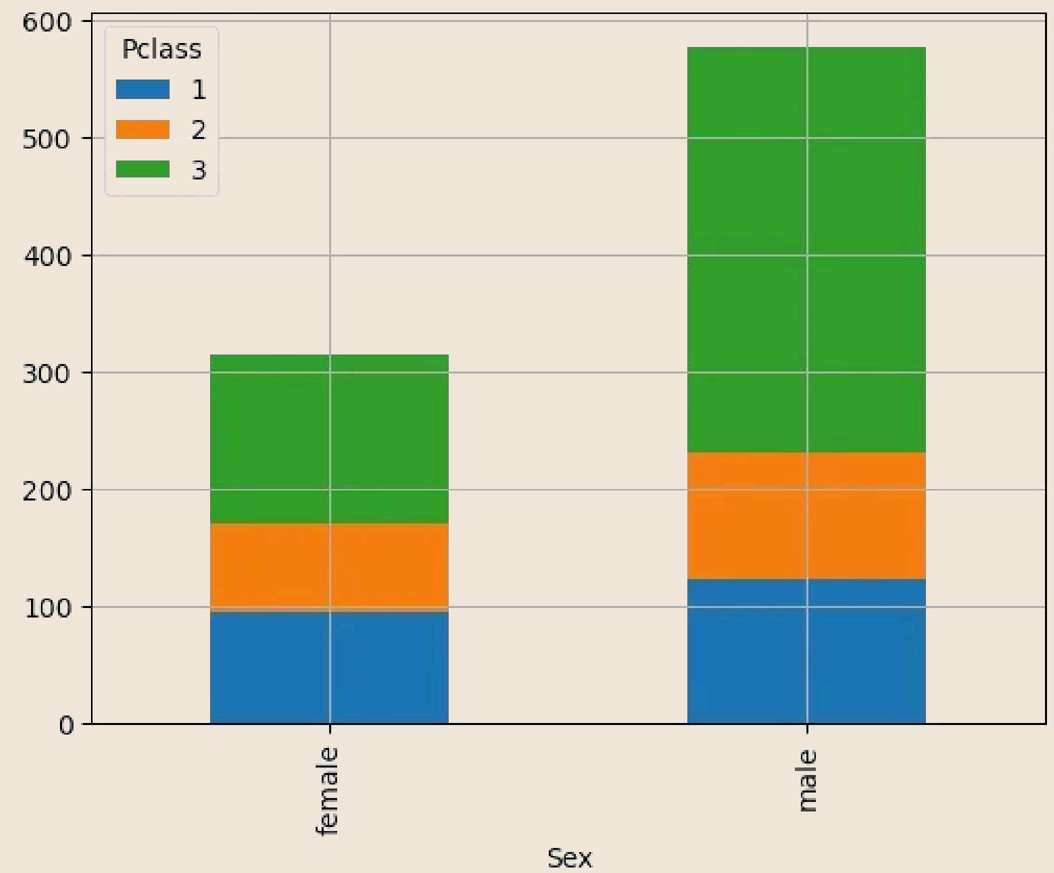
```
pd.crosstab(titanic['Sex'], titanic['Pclass']).plot.bar()
plt.grid()
plt.show()
```

Graphical representation of absolute counts through the contingency table.




```
pd.crosstab(titanic['Sex'], titanic['Pclass']).plot.bar(stacked=True)
plt.grid()
plt.show()
```

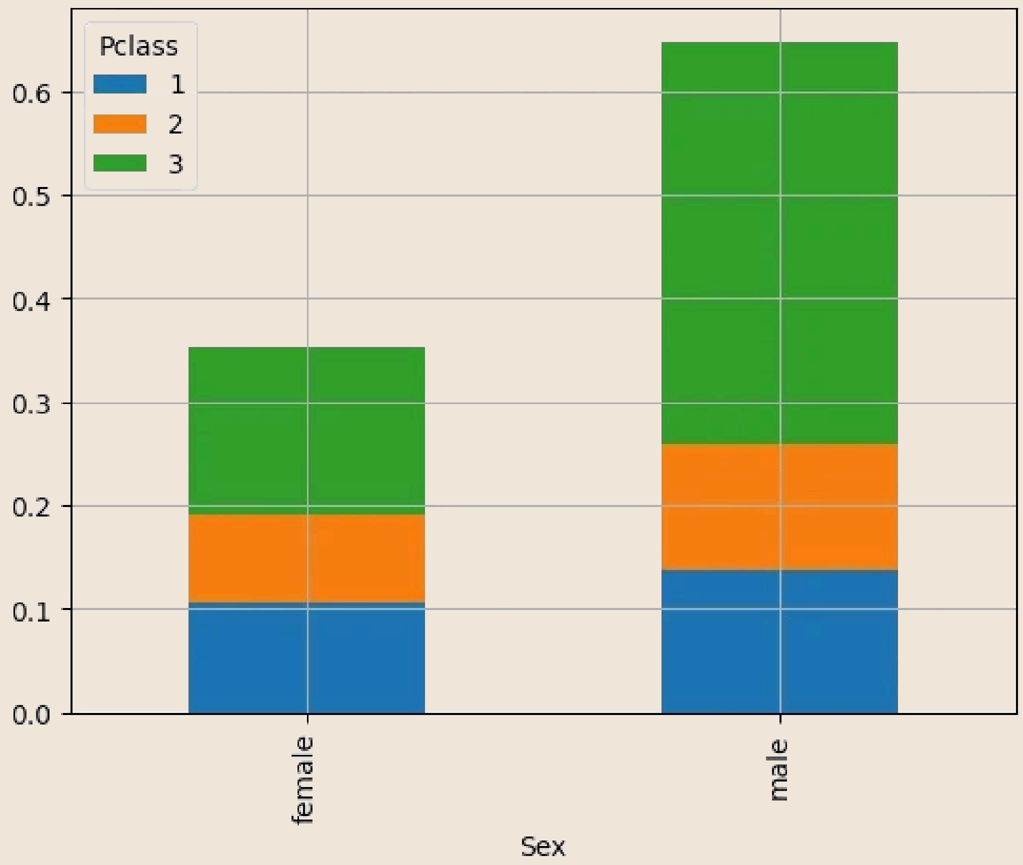
Stacked version of the previous plot.



Graphical Representations — cont.

```
pd.crosstab(titanic['Sex'], titanic['Pclass'], normalize=True).plot.bar(stacked=True)
plt.grid()
plt.show()
```

Normalized version: it shows the proportions of first second and third class across males and females.



Sum Rule (Marginal Probability)

We have seen how to compute **marginal (univariate) probabilities** from the contingency table. In general, we can **compute marginal probabilities form joint probabilities** (i.e., we don't need to have the non-normalized frequency counts of the contingency table). Let us consider the general contingency table:

	Y=y_1	**Y=y_2**	**...**	**Y=y_l**	**Total**
X=x_1	n_{11}	n_{12}	...	n_{1l}	n_{1+}
X=x_2	n_{21}	n_{22}	...	n_{2l}	n_{2+}
...	...	...	...	...	...
X=x_k	n_{k1}	n_{k2}	...	n_{kl}	n_{k+}
Total	n_{+1}	n_{+2}	...	n_{+l}	n

We can compute joint probability using the frequentist approach as follows:

$$P(X=x_i, Y=y_j) = \frac{n_{ij}}{n}$$

Hence, we can compute the marginal probability as follows:

$$P(X=x_i) = \frac{n_{i+}}{n} = \frac{\sum_j n_{ij}}{n} = \sum_j \frac{n_{ij}}{n} = \sum_j P(X=x_i, Y=y_j)$$

This leads us to the definition of the **Sum Rule** of probability:

$$P(X=x) = \sum_y P(X=x, Y=y)$$

The act of computing P(X) from P(X,Y) is also known as marginalization.

Conditional Probability

Probability of an event **given** that another event occurred

$$P(X=x|Y=y)=\frac{P(X=x,Y=y)}{P(Y=y)}$$

Read as: "probability of X equals x, **given that** Y equals y"

Example: P(S=male | C=1) = "probability of male given first class"

Let's consider the generic contingency table:

	Y=y_1	**Y=y_2**	**...**	**Y=y_l**	**Total**
X=x_1	n_{11}	n_{12}	...	n_{1l}	n_{1+}
X=x_2	n_{21}	n_{22}	...	n_{2l}	n_{2+}
...	...	...	...	...	...
X=x_k	n_{k1}	n_{k2}	...	n_{kl}	n_{k+}
Total	n_{+1}	n_{+2}	...	n_{+l}	n

We can compute the conditional probability with the frequentist approach:

$$P(X=x_i|Y=y_j)=\frac{\#\,cases\ in\ which\ X=x_i\ and\ Y=y_i}{\#\,cases\ in\ which\ Y=y_j}=\frac{n_{ij}}{n_{+j}}$$

Multiplying both terms by $1=\frac{n}{n}$, we obtain:

$$P(X=x_i|Y=y_j)=\frac{n_{ij}}{n}\frac{n}{n_{+j}}=\frac{\frac{n_{ij}}{n}}{\frac{n_{+j}}{n}}=\frac{P(X=x_i,Y=y_j)}{P(Y=y_j)}$$

The conditional probability is defined only when P(Y = y) > 0, that is, we cannot define a probability conditioned on an event that never happens. It should be noted that, in general:

$$P(X|Y)\neq P(X)$$

Product Rule (Factorisation)

We can write the definition of conditional probability as follows:

$$P(X = x, Y = y) = P(X = x | Y = y) \cdot P(Y = y)$$

This is known as **product rule** or **factorization** as it allows to factorize a joint probability into the product of a conditional or marginal.

☐ The product rule allows us to compute joint probabilities without building large tables

Computing conditional probabilities is often easier—we just **restrict** observations to those satisfying the condition

Computing Conditional Probabilities

Let's consider this contingency table:

	First Class	Second Class	Third Class	All
male	3	5	7	15
female	3	3	4	10
All	6	8	11	25

Note that we only need the first column to compute the conditional probability:

$$P(S|C=1)$$

Indeed:

- $P(S=male|C=1)=\frac{3}{6}$
- $P(S=female|C=1)=\frac{3}{6}$

Computing Conditional Probabilities in Python

```
# normalize=0 indicates to condition on the first variable
pd.crosstab(titanic['Sex'], titanic['Pclass'], normalize=0, margins=True)
```

Pclass	1	2	3
Sex			
female	0.299363	0.242038	0.458599
male	0.211438	0.187175	0.601386
All	0.242424	0.206510	0.551066

From the table above, for example, we can infer:

- $f(Pclass = 1|Sex = female) = 0.290363$
- $f(Pclass = 2|Sex = female) = 0.242038$
- $f(Pclass = 3|Sex = female) = 0.458599$

We can obtain the complementary perspective by conditioning on class instead:

```
# normalize=1 indicates conditioning on the first variable
pd.crosstab(titanic['Sex'], titanic['Pclass'], normalize=1, margins=True)
```

Pclass	1	2	3	All
Sex				
female	0.435185	0.413043	0.293279	0.352413
male	0.564815	0.586957	0.706721	0.647587

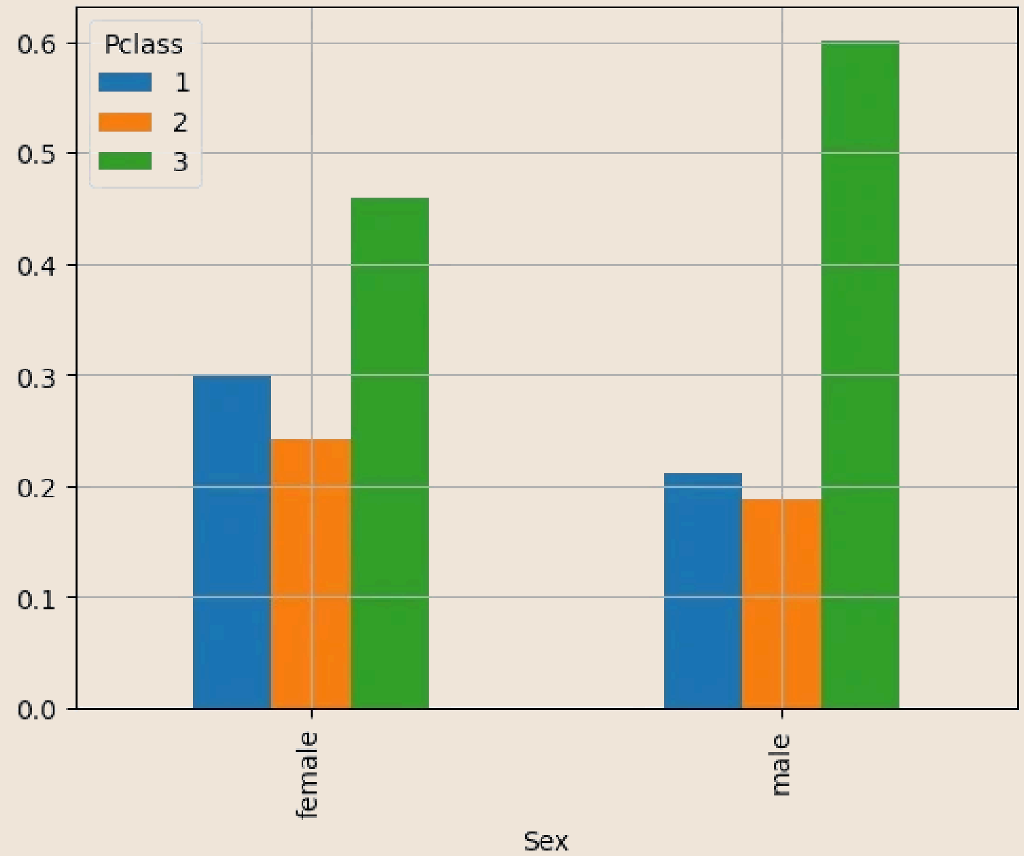
In this case, each column will be a probability distribution. For example:

- $P(Sex = female|Pclass = 1) = 0.435185$
- $P(Sex = male|Pclass = 1) = 0.564815$

We note that the proportion between men and women changes in the three classes, and in particular in the third class there are many more men than women.

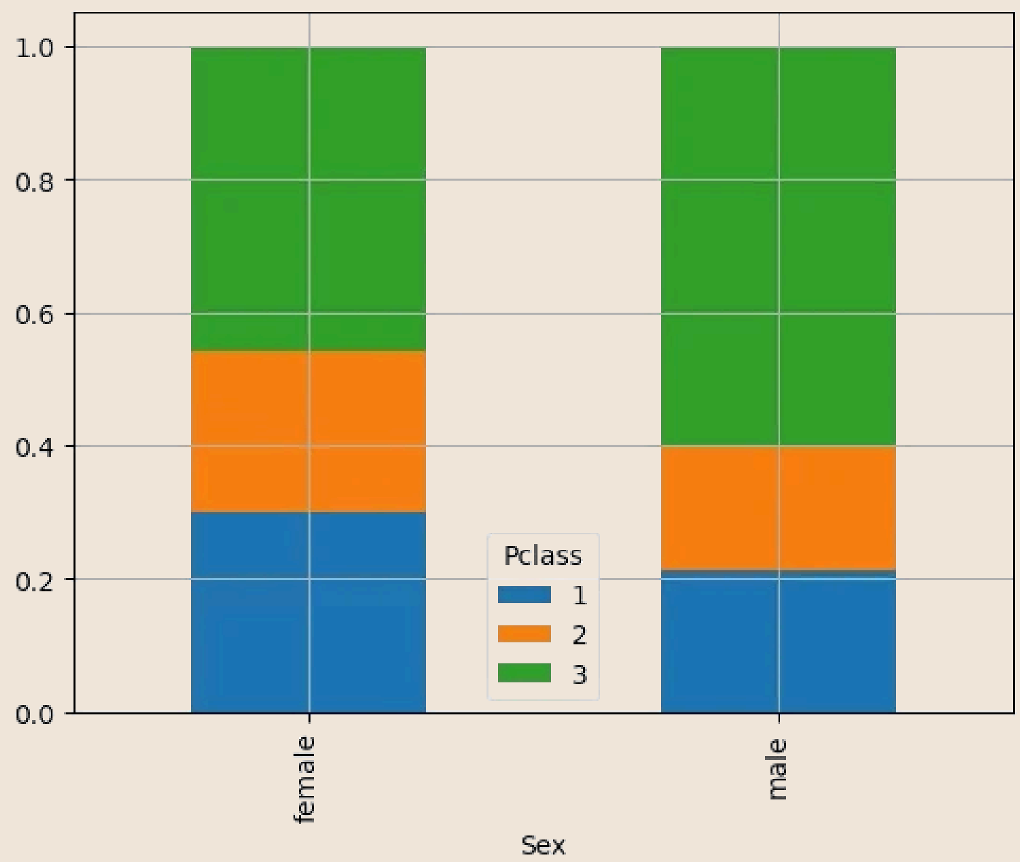
Visualizations from Conditional Probabilities

```
pd.crosstab(titanic['Sex'], titanic['Pclass'], normalize=0).plot.bar()
plt.grid()
plt.show()
```



Distributions of classes for male and female cases. Note that these are two separate probability distributions, each summing to one.

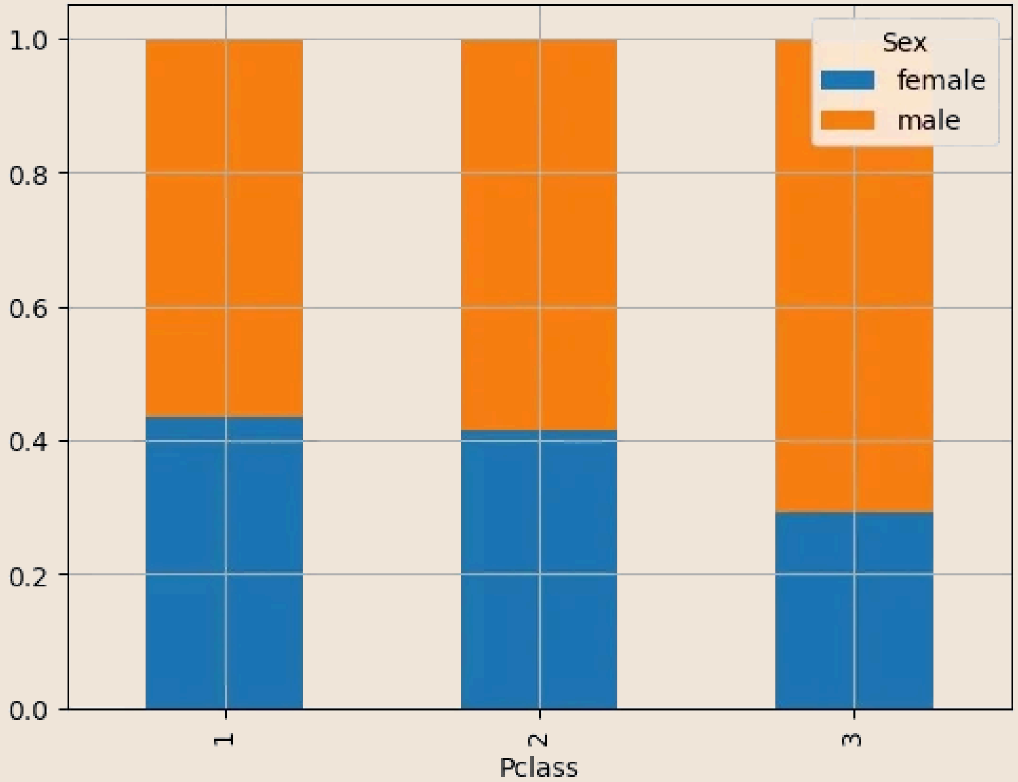
```
pd.crosstab(titanic['Sex'], titanic['Pclass'], normalize=0).plot.bar(stacked=True)
plt.grid()
plt.show()
```



Stacked version of the plot below. Highlights the difference in distributions of class across sexes.

Visualizations from Conditional Probabilities — cont.

```python
# normalize=1 indicates conditioning on the first variable
pd.crosstab(titanic['Sex'], titanic['Pclass'], normalize=1).T.plot.bar(stacked=True)
plt.grid()
plt.show()
```



Other view: now we have three conditional distributions, one per class, and each summing to one.

Chain Rule of Conditional Probability

The Chain Rule expresses the joint probability of multiple random variables as a product of conditional probabilities, simplifying complex probabilistic models.

For two variables:

$$P(A, B) = P(A|B)P(B)$$

For multiple variables:

$$P(X_1, \ldots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \ldots P(X_n|X_1, \ldots, X_{n-1})$$

More generally:

$$P(X_1, \ldots, X_n) = P(X_1) \prod_{i=2}^{n} P(X_i|X_1, \ldots, X_{i-1})$$

Bayes' Theorem

Bayes' Theorem updates the probability of an event based on new evidence, a key concept in statistical inference and AI.

The theorem is expressed as:

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

Where:

- **P(A|B)**: Posterior probability (updated belief in A after observing B).
- **P(B|A)**: Likelihood (probability of observing B if A is true).
- **P(A)**: Prior probability (initial belief in A before observing B).
- **P(B)**: Evidence (total probability of observing B).

Bayesian Probability Example

Scenario: You want to estimate the probability that your friend has COVID-19, given that they currently have a fever.

- **Prior Probability P(COVID):** Assume that, in the general population, the probability of someone having COVID is **0.5** (1 in 2 people).
- **Likelihood P(fever|COVID):** The probability of having a fever given that one has COVID is **1/3** (1 in 3 COVID patients has a fever).
- **Evidence P(fever):** The overall probability of someone having a fever (regardless of COVID status) in the general population is **1/5** (1 in 5 people).

We want to find the **Posterior Probability P(COVID|fever)** – the probability that your friend has COVID, given they have a fever.

Using Bayes' Theorem:

$$P(COVID|fever) = \frac{P(fever|COVID) \cdot P(COVID)}{P(fever)}$$

Plugging in the values:

$$P(COVID|fever) = \frac{(1/3) \cdot (0.5)}{(1/5)} = \frac{0.1666}{0.2} = 0.8333 \approx 0.83$$

Interpretation: Our prior belief that your friend has COVID was 50%. After observing the new evidence (your friend has a fever), our updated belief (posterior probability) that they have COVID increases significantly to approximately **83%.**

Independence of Variables

Two random variables, X and Y, are considered **independent** if the occurrence of one event does not influence the probability of the other.

Mathematically, independence is defined in two equivalent ways:

- **Conditional Probability:** The probability of X given Y is simply the probability of X.

$$P(X = x|Y = y) = P(X = x)$$

- **Joint Probability:** The joint probability of X and Y is the product of their individual probabilities.

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$$

Examples

Intuitively, two variables are independent if the values of one of them do not affect the values of the other one:

- Weight and height of a person are **not independent**. Indeed, taller people are usually heavier.
- Height and richness are **independent**, as the richness does not depend on the height of a person.

Probability Manipulation Examples

Let's consider a real probability manipulation example using the Titanic dataset with variables Pclass and Survived.

Random Variables:

- **C:** Passenger class {1, 2, 3}
- **S:** Survival outcome {0 = died, 1 = survived}

Questions we'll answer using observed proportions:

- What is the overall probability that a passenger survived?
- What is the probability that a passenger was in 1st class and they survived?
- What is the probability that a passenger survived, given that they were in 3rd class?

Conclusions and Next Steps

We Have Explored:

- Definition of probability
- Joint Probabilities
- Marginal Probabilities
- Conditional Probability
- Sum Rule
- Product Rule
- Chain Rule
- Bayes' Theorem

> In next lectures, we will look at association between variables

References

- Parts of chapter 1 of [1];
- Most of chapter 3 of [2];
- Parts of chapters 5-7 of [3].

[1] Bishop, Christopher M. *Pattern recognition and machine learning.* springer, 2006. **Link**

[2] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning.* MIT press, 2016. **Link**

[3] Heumann, Christian, and Michael Schomaker Shalabh. Introduction to statistics and data analysis. Springer International Publishing Switzerland, 2016.

- Lecture notes: **Chapter 3 — Probability for data analysis**