Statistics for Astronomers: Lecture 03, 2019.02.14

Prof. Sundar Srinivasan

IRyA/UNAM

Recall: Computing probability: three interpretations

- **Classical interpretation**: can be computed if all outcomes are assumed equally likely.

  We just used this interpretation to compute the probability in the previous slide.
- **Frequentist interpretation**: perform an infinite sequence of experiments, find relative frequency of favourable outcomes.

  Toss the two coins N times ($N \gg 1$) and count the number of times M that we get two tails. The required probability is then $\frac{M}{N}$.
- **Bayesian interpretation**: use prior knowledge of the parameters of the problem, perform experiments, and update the priors to get posterior probabilities.

  Select your priors (*e.g.*, are the coins known to be fair from experience? "Roberto performed the experiment 500 times yesterday and only got two tails 20 times!"). Perform an experiment. Combine the resulting outcome with the prior and predict the probability of getting two tails on future trials.

Recall: Classical (naïve) interpretation of probability

"The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible..."
– Laplace (1812).

Principle of Indifference

If N events are mutually exclusive and collectively exhaustive,

(1) they are equally likely and (2) the probability of any one occurring is $\frac{1}{N}$.

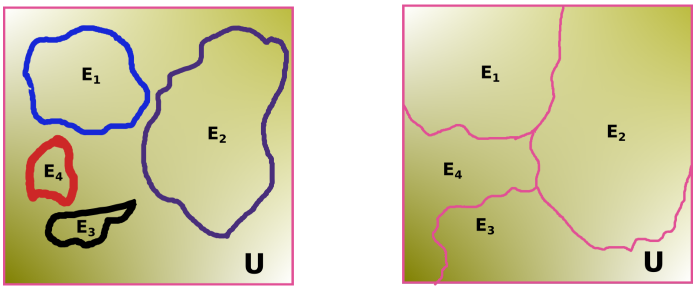


Figure: Case (1): E1, E2, E3, and E4 are mutually exclusive but not collectively exhaustive. Case (2): the events are now also collectively exhaustive.

Recall: Axioms of Probability (Kolmogorov, 1933)

1. The probability that an event has occurred is always a non-negative real number.
   $\forall E \in \mathscr{F}, P(E) \in \mathbb{R} \text{ and } P(E) \geq 0$
   In particular, $P(\varnothing) = 0$. (At least one event in Ω occurs.)
2. **Unitarity:** The probability that at least one event in the sample space will occur is unity.
   $P(\Omega) = 1$
3. **Countable additivity:** The probability that at least one event among a set of (pairwise) disjoint events occurs is the sum of the probabilities of each of those events occurring.
   Given A_j $(j = 1, \ldots)$ such that $A_j \cap A_k = \varnothing$ for $j \neq k$,
   $$P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j), \text{ if } A_j \cap A_k = \varnothing \text{ for } j \neq k$$

Recall: Conditional probability and indepdendence

Definition (Conditional probability)

The probability that an event A occurs, given that another event B has already occurred. Representation: $P(A|B)$ ("probability of A given B").

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

$$P(A \text{ occurs, given that } B \text{ already occurred}) = \frac{P(A \text{ and } B \text{ both occur})}{P(B \text{ occurs})}$$

Definition (Independence)

Two events A and B are said to be independent ("$A \perp B$") if the occurrence of one does not affect the probability of occurrence of the other.

$A \perp B \Rightarrow P(A \cap B) = P(A) \times P(B)$.

Mutually exclusive events are not independent.

Recall: Conditionality and Marginalisation

Using conditional probabilities, $P(A) = P(A|B) \times P(B) + P(A|B^c) \times P(B^c)$.
$P(A)$ is obtained by "mariginalising over B".

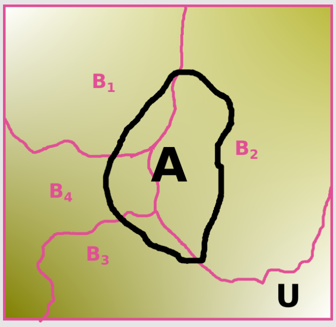
Generalisation: Law of Total Probability

(Connects conditional probabilities to marginal probability)

Given N pairwise disjoint and collectively exhaustive events B_i ($i = 1, 2, ..., N$), the probability of occurrence of an event A is given as the weighted average of the conditional probabilities $P(A|B_i)$, with weights $P(B_i)$:

$$P(A) = \sum_{i=1}^{N} P(A \cap B_i) = \sum_{i=1}^{N} P(A|B_i) \times P(B_i)$$

$P(A)$ is then the probability of A marginalised over the events B_i.



Recall: Bayes' Theorem

Definition (Bayes' Theorem)

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

Under the Bayesian Interpretation of probability, this is read as

Updated deg. of belief in A = Support for A from evidence B × Original deg. of belief in A.

or

$$\text{Posterior prob. of } A \text{ given evidence } B = \frac{\text{Cond. prob. of } B \text{ given } A}{\text{Marginal prob. of } B} \times \text{Prior prob. of } A.$$

or

$$\text{Posterior prob. of } A \text{ given evidence } B = \frac{\text{Likelihood of } A \text{ given } B}{\text{Evidence } B} \times \text{Prior prob. of } A.$$

We can use the Law of Total Probability to convert the marginal probability into conditional probabilities.

Bayes' Theorem implies that, in general, $P(A|B) \neq P(B|A)$.
Many logical fallacies arise from misunderstandings of this relationship.

The Prosecutor's Fallacy

"The probability of obtaining evidence against you, given that you are innocent, is very small. Therefore, the fact that evidence was obtained against you means that you are probably guilty."

This argument does not take any prior information into account, it is completely evidence-based.
The prior probability that any person is innocent is, in general, high.

Let I = "The person is innocent" and E = "Evidence was found against the person."
The Fallacy can then be stated as $P(I|E) = P(E|I)$.

Bayes' Theorem instead states that $P(I|E) = P(E|I) \times P(I)/P(E)$.
In general, $P(I)$ is quite high and $P(E)$ quite low, so that $P(I|E)$ may be greater than $P(E|I)$.

Read about the Sally Clark Case.

Bayesian inference versus frequentist inference

Bayesian inference relies on assuming prior knowledge of the hypotheses/parameters of interest. In the Bayesian interpretation, probability is a *degree of belief*. In this sense, Bayesian probabilities are subjective.

Frequentist inference is, by contrast, considered objective as it does not incorporate/assume priors for the parameters that are the subject of inference.

Random variables and probability distributions

Random variables

Random: Uncertain, no "pattern" can be detected.

Randomness: A measure of uncertainty of the outcome of an experiment. Some sources of "true" randomness – initial conditions of the experiment (*e.g.*, throwing dice, chaos) and environmental effects (*e.g.*, Brownian Motion, dark current).

Random variable: A function that assigns a numerical value to each distinct outcome. Allows us to compute probabilities.

Random process: A sequence of random variables whose outcomes don't follow a pattern. Their evolution can, however, be described probabilistically.
Each observation results in a random variable associated with the process.
The collection of random variables in such a process have two attributes: an index drawn from an index set, and a numerical value drawn from a state space.
Each outcome is mapped to a unique element of the index set, and each unique outcome is mapped to a unique value in the state space.

Random variable: A function that maps the sample space to the state space; $X : \Omega \to S$.

Probability distribution: A function that maps a random variable to a real number; $p : X \to \mathbb{R}$.

Illustration: index sets and state spaces

Tossing a fair coin $N = 5$ times: A sequence of $N = 5$ random variables.
Number these trials from 1 to 5 $\Rightarrow$ index set $= \{1, 2, 3, 4, 5\} \subset \mathbb{Z}$.
Sample space for each toss, $\Omega = \{T, H\}$.
Map outcomes to two distinct values: $T \to 0$, $H \to 1 \Rightarrow$ state space $= \{0, 1\} \subset \mathbb{Z}$.
$E =$ "First toss results in heads" $\equiv X_1 = 1 \Rightarrow P(E) = P(X_1 = 1) = \dfrac{1}{2}$.

Choosing three balls at random without replacement from a box containing 2 red balls and 8 orange balls:
3 trials $\Rightarrow$ Index set $= \{1, 2, 3\} \subset \mathbb{Z}$.
$\Omega = \{\mathrm{Red}, \mathrm{Orange}\}$. Choose state space $= \{0, 1\} \subset \mathbb{Z}$.
$E =$ "Third ball picked is red" $\equiv X_3 = 0 \Rightarrow P(E) = P(X_3 = 0) = \dfrac{3}{7}$.

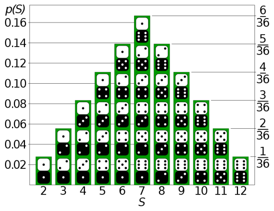
Actual departure time of AA 1066 (MEX → DFW, scheduled: 6:00 AM) over 10 trips:
Index set $= \{1, ..., 10\} \subset \mathbb{Z}$. Could also choose $\subset \mathbb{R}$!.
$\Omega \subset \mathbb{R}$ (uncountably infinite values!) State space $\subset \mathbb{R}$.
$E =$ "On the 2$^{\mathrm{nd}}$ day, it departed at 8:22 AM". $P(E) = ?$

Illustration: probability distributions

Choosing three balls at random without replacement from a box containing 2 red balls and 8 orange balls (discrete distribution):

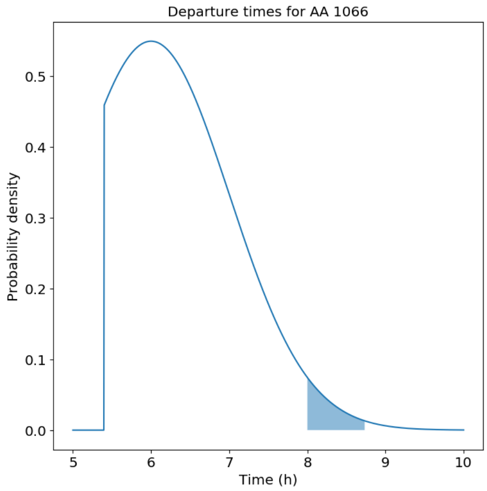
Distribution of probabilities for colour of third ball: $P(X_3 = 0) = \frac{3}{7}, P(X_3 = 1) = \frac{4}{7}$.

Sum of numbers displayed on two dice after one throw (discrete distribution):
State space $= \{2, 3, ..., 11, 12\}$
$P(4 \leq X \leq 7) = \frac{18}{36} = \frac{1}{2}$.



(Tim Stellmach/Public Domain)

Departure time of AA 1066 on the 2$^{\rm nd}$ of 10 days (continuous distribution):



Code for plot available here.

Shaded region $= P(8{:}00\ \mathrm{AM} < T < 8{:}44\ \mathrm{AM})$.

Discrete and continuous probability distributions

A discrete distribution is also called a probability mass function (PMF).
The probability of the random variable spanning a range of values is the sum of the PMFs for each value. In the two-dice example from the previous slide, $P(4 \leq X \leq 7) = \sum_{x=4}^{7} P(X = x)$.
The PMF for a single value of the random variable, $X = x$, is equal to the probability that $X = x$.

Continuous distributions are called probability density functions (PDFs).
The probability of the random variable spanning a range of values is the integral of the PDF over the range. For example, if $p(X)$ is the PDF,

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x) dx.$$

While the PDF for a single value of the random variable, $X = x$, may be nonzero, the probability that $X = x$ is zero ($x_1 = x_2$ in the integral above).

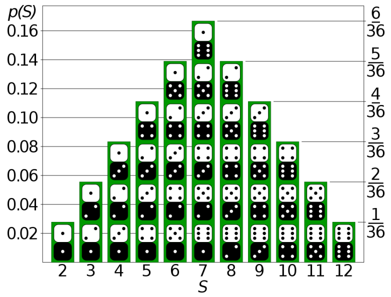
Notation: I will use P for total probability or PMF, and p for the PDF.
Note: I won't abbreviate "probability distribution function", so that "PDF" is unambiguous.

Populations and samples

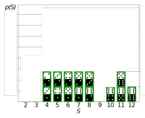
If an experiment results in a random variable X whose probability distribution is $P_{X3pt}(x)$ (discrete) or $p(X)$, we say that X is drawn from the PMF/PDF: $X \sim P(X)$ or $X \sim p(X)$.

Population: the underlying probability distribution.
Sample: the results of a finite number of experiments/draws from the population (a subset).

Due to the fact that we perform the experiment a finite number of times, the sample may not be able to faithfully reproduce the population – statistical quantities derived from the sample are only guesses at the quantities related to the population. Consider the two-dice example from before:



(Tim Stellmach/Public Domain)
Population mean: 7.00, population variance: 5.83



Sample of outcomes obtained from rolling two dice 14 times.
Sample mean: 7.43, sample variance (discussed later): 6.67

Cumulative distribution function (CDF)

Definition (Cumulative distribution function)

A function $F_X(x)$ of a random variable X such that $F_X(x)$ is the probability that $X \leq x$.
For a discrete PMF: $F_X(x) = P(X \leq x) = \sum\limits_{x_i \leq x} P(X = x_i)$.

For a continuous PDF: $F_X(x) = P(X \leq x) = \int\limits_{t=-\infty}^{t=x} p(t)dt$.

From this definition, the probability of the variable ranging between two values a and b is

$P(a < X \leq b) = F_X(x = b) - F_X(x = a)$. For a PDF, this is also equal to $\int\limits_{t=a}^{t=b} p(t)dt$.

The CDF is a monotonically increasing function. For a discrete random variable, it is constant in between values.

For the continuous case, the PDF is the derivative of the CDF w.r.t. x.

Cumulative distribution function (contd.)

Definition (Quantile function)

The inverse of the CDF, a function $Q(p)$ that returns the value of x such that $F_X(X \leq x) = p$.
e.g., $Q(p = 0.5)$ is the median (equal "mass" on either side of $x = Q(0.5)$).
$Q(p = 0.25)$ and $Q(p = 0.75)$ are the first and third quartiles.

Definition (Independent and identically distributed variables)

Two random variables X and Y are said to be iid if and only if they are mutually independent and drawn from the same distribution:

$$F_{X,Y}(x,y) = F_X(x) \times F_Y(y)$$
$$F_X(x) = F_Y(x)$$

Expectation value

Definition (Expectation value)

The expectation value of any function $g(X)$ of a random variable X, represented by $E[g(X)]$, is the weighted average of $g(X)$, with the weights being the associated probabilities:

$E[g(X)] = \sum_{i=1}^{N} g(x_i)\ P(X = x_i)$ (discrete case)

$E[g(X)] = \int_{t=-\infty}^{t=\infty} g(x)\ p(x)\ dx$ (continuous case)

The expectation value is a linear operator, so that, for any two functions $g(X)$ and $h(X)$ and scalars α and β,
$E[\alpha g(X) + \beta h(X)] = \alpha E[g(X)] + \beta E[h(X)]$.
Also, if $X \perp Y$, then $E[XY] = E[X]E[Y]$.

Some common expectation values incude the mean $E[X]$ and the variance $Var[X] \equiv E[(X - E(X))^2]$ (equal to the square of the standard deviation).
Why the square? What is $E[X - E(X)]$?
The expression for the variance can be simplified: $Var[X] = E[X^2] - (E[X])^2$.

Note: When no other function is specified, "expectation value" refers to the mean, $E(X)$.

Discrete probability distributions

Bernoulli Distribution

A Bernoulli random variable takes one of only two values: 1 and 0, with probabilities p and $1-p$ respectively. It is the result of an experiment that asks a single yes-no question.
The variable therefore has state space $S = \{1, 0\}$, with an associated probability distribution given by

Definition (Bernoulli Distribution)

$P(X=1) = p$ and $P(X=0) = 1-p$ (Bernoulli Distribution). We can abbreviate this:
$P(X=x) = p^x(1-p)^{1-x}\ \mathbb{I}_{x\in\{0,1\}}(x)$, with $\mathbb{I}(x)$ the Indicator (or Heaviside) function.

Example of a Bernoulli random variable: outcome of tossing a single (not necessarily fair) coin.

Mean: $E[X] = 1 \times P(X=1) + 0 \times P(X=0) = 1 \times p + 0 \times (1-p) = p$
Variance: First, $E[X^2] = 1^2 \times P(X=1) + 0^2 \times P(X=0) = 1^2 \times p + 0^2 \times (1-p) = p$
$\Rightarrow Var[X] = E[X^2] - (E[X])^2 = p - p^2 = p(1-p)$