



SCHOOL OF MATHEMATICS
AND PHYSICS

MTH1005 PROBABILITY AND STATISTICS

Semester B

(13/2/2024)

Danilo Roccatano

Office:

INB 3323

Email:

droccatano@lincoln.ac.uk

CONDITIONAL PROBABILITY AND INDEPENDENCE



Learning outcomes:

- Conditional probability
- Definition of independence of events
- Bayes' Theorem
- Tree and Venn diagrams
- Law of total probability

PROBABILITY FUNCTION

In the previous lecture, we defined the probability function (sometimes also called a probability distribution function) for a sample space, S .

The sample space lists all possible outcomes of an experiment. For instance, for a single roll of a 6-sided die, a possible sample space is $S = \{1, 2, 3, 4, 5, 6\}$.

It is a function that, if you feed it an outcome of an experiment, produces a probability x ,

$$P(i \in S) = P(\{i\}) = 0 \leq x \leq 1$$

Using this function, we can calculate the probability of an event, E

$$P(E) = \sum_{i \in E} p(i)$$

i.e. we add together the probabilities of all the outcomes that are in the event, E .

PROBABILITY FUNCTION



EXAMPLE

For instance, the

'getting an even number larger than 3 when rolling a single dice' =

$$A = \{2, 4, 6\} \cap \{4, 5, 6\} = \{4, 6\}$$

is

$$P(A) = \sum_{i \in A} p(i) = p(4) + p(6) = \frac{2}{6}$$

CONDITIONAL PROBABILITY



The concept of conditional probability enables us to compute probabilities in situations where we possess only partial information about the system or to reevaluate probabilities upon receiving new information.

In simpler terms, conditional probability refers to the probability of an event E transpiring given that another event F has already occurred.

This is represented mathematically as

$$P(E | F)$$

CONDITIONAL PROBABILITY

Example: Consider rolling two dice. We have a sample space

$$S = \{(i, j), i = 1, 2, 3, 4, 5, 6, j = 1, 2, 3, 4, 5, 6\}$$

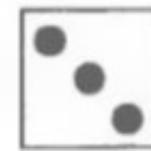
where say the outcome (i, j) is the first die lands with i dots up and the second die with j .

We assume the dice are fair - we assign a probability (distribution) function of

$$P(\{(i, j)\}) = p(i, j) = \frac{1}{36}$$

to all outcomes.

Suppose the first die comes down a three,

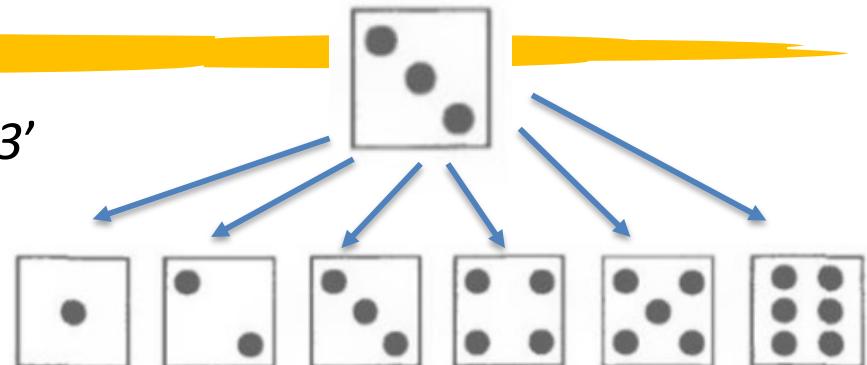


what is the probability that the sum of the two dice equals eight?

CONDITIONAL PROBABILITY

Define $F = \text{'the first die comes down showing a 3'}$

$$F = \{(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6)\}.$$



This extra information, allows us to cut down the number of possible outcomes. Effectively only 6 possible outcomes can now occur. Each of these remaining outcomes is still equally likely (what do you think?).

So, to get the probability we are after we need to find the proportion of the remaining 6 outcomes that are in

$$E = \text{'the sum of the dice} = 8\text{'} = \{(2, 6), (3, 5), (4, 2), (5, 3), (6, 2)\}.$$

The only outcome that satisfies the criteria is $\{(3, 5)\}$ and once we know that F has occurred, it has a one in 6 chance that $\{(3, 5)\}$ will occur.

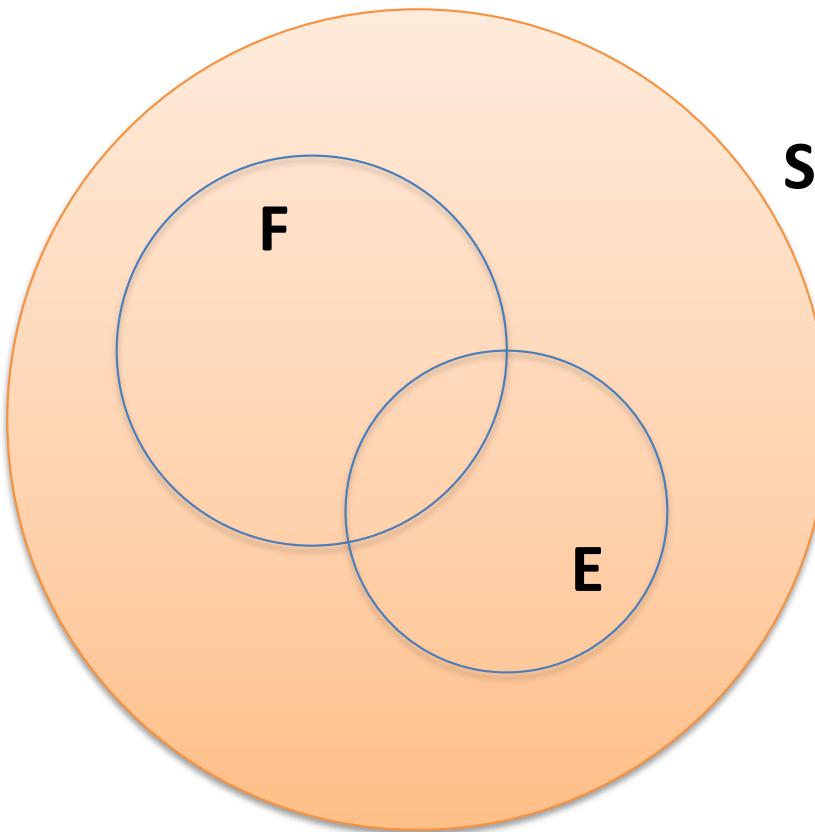
$$P(E | F) = \frac{1}{6}$$

Note also that $\{(3, 5)\} = E \cap F$

CONDITIONAL PROBABILITY

The probability of the outcome (3,5) before rolling the first die was $1/36$. Following the result of the first die showing a 3, we have asserted that the probability should be $1/6$.

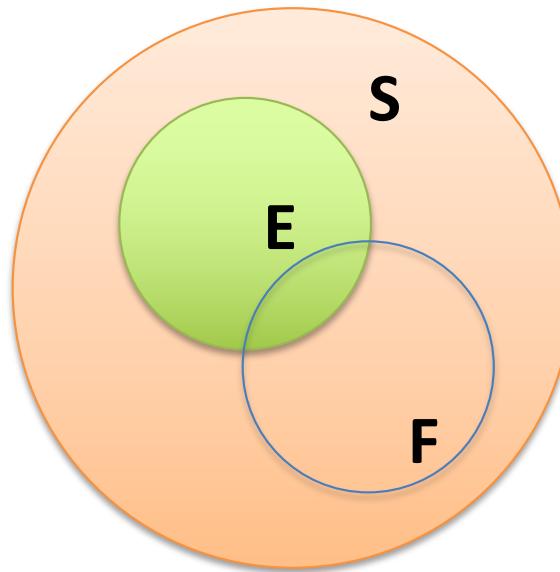
Now, let's visualize this scenario using a Venn diagram:



where S is the sample space, and we have labelled two events E and F .

CONDITIONAL PROBABILITY

Now the probability of E occurring is the probability of the outcomes in E .



If we are dealing with finite sets with equal likelihood, this probability is

$$\frac{n(E)}{n(S)}$$

where $n(E)$ is the number of ways that the event E can occur, and $n(S)$ the same for S .

CONDITIONAL PROBABILITY

If, however, we are seeking to calculate $P(E | F)$, then the ratio of $n(E)$ to $n(S)$ becomes irrelevant:

Given that event F has already occurred, our focus shifts to determining the number of occurrences where both E and F happen simultaneously. However, in this context, our effective sample space is constrained to just F .

This gives

$$P(E | F) = \frac{P(E \cap F)}{P(F)}$$

For discrete equal-likelihood sample spaces, this is:

$$P(E | F) = \frac{n(E \cap F)}{n(S)n(F)} = \frac{n(E \cap F)}{n(F)}$$

CONDITIONAL PROBABILITY

If we look back to our dice example, this is exactly what we did.

Example revisited. we started from

$$S = \{(i, j), i = 1, 2, 3, 4, 5, 6, j = 1, 2, 3, 4, 5, 6\},$$

and asked, "Suppose the first die comes down with a three, what is the probability that the sum of the two dice equals eight?"

In the language we've just used, our event E is

$$E = \{(2, 6), (6, 2), (3, 5), (5, 3), (4, 4)\},$$

the event

$$F = \{(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6)\}.$$

So

$$E \cap F = \{(3, 5)\}.$$

CONDITIONAL PROBABILITY



Looking at the size of these events (all individual outcomes are equally likely), we have

$$P(E | F) = \frac{n(E \cap F)}{n(S)n(F)} = \frac{n(E \cap F)}{n(F)} = \frac{1}{6}$$

CONDITIONAL PROBABILITY

Definition: the conditional probability of E given F is

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

combined with the standard properties of probabilities this implies

- $P(\bar{E}|F) = 1 - \frac{P(E \cap F)}{P(F)}$
- if E and F are mutually exclusive, then $P(E|F) = 0$ because $E \cap F = \emptyset$

MULTIPLICATIVE RULE OF PROBABILITIES

The expression for conditional probabilities implies that

$$P(E \cap F) = P(E | F)P(F) \quad (1)$$

In simple terms, this equation states that the probability of both events E and F occurring is equal to the probability of event F occurring multiplied by the probability of event E occurring given that F has already occurred.

This can be generalized to the case of many events E_i

$$\begin{aligned} P(E_1 \cap E_2 \cap E_3 \cap \dots \cap E_n) &= P(E_1) \cdot P(E_2 | E_1) \cdot P(E_3 | E_1 \cap E_2) \\ &\dots P(E_n | E_1 \cap \dots \cap E_{n-1}) \end{aligned}$$

MULTIPLICATIVE RULE OF PROBABILITIES

To demonstrate this relation, let's substitute the definition of conditional probability into the right side of the generalized equation:

$$\begin{aligned} P(E_1 \cap E_2 \cap E_3 \cap \dots \cap E_n) &= P(E_1) \cdot P(E_2 | E_1) \cdot P(E_3 | E_1 \cap E_2) \\ &\dots P(E_n | E_1 \cap \dots \cap E_{n-1}) \end{aligned}$$

Using the definition of conditional probability, $P(E \cap F) = P(E | F) P(F)$, we can rewrite each conditional probability term in the equation above:

$$P(E_1) \frac{P(E_1 \cap E_2)}{P(E_1)} \cdot \frac{P(E_1 \cap E_2 \cap E_3)}{P(E_1 \cap E_2)} \dots \frac{P(E_1 \cap E_2 \dots \cap E_n)}{P(E_1 \cap E_2 \dots \cap E_{n-1})}$$

each internal term of the numerator cancels with the next term in the denominator.

MULTIPLICATIVE RULE OF PROBABILITIES



EXAMPLE: Let's revisit the scenario of selecting balls from a bowl, as discussed in lecture 1, where there were 6 black balls and 5 white ones.

How many ways exist to select all the balls from the bowl?

In this case, the number of ways can be calculated using permutations, resulting in $N=11!$ (11 factorial).

To elaborate, we can approach the calculation as follows:

Since each permutation is equally likely, we can compute the probability of obtaining a specific selection, which would be $1/N$.

Therefore, we determine N by taking the inverse of the probability.

MULTIPLICATIVE RULE OF PROBABILITIES

If we focus on the specific selection:

$$\{(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)\}$$

where each ball is numbered accordingly, we can define

$E_n = \text{'the } n\text{th ball is ball } n\text{'}$

$$\begin{aligned} P(E_1 \cap E_2 \cap E_3 \cap \dots \cap E_n) &= P(E_1) \cdot P(E_2 | E_1) \cdot P(E_3 | E_1 \cap E_2) \\ &\quad \dots P(E_n | E_1 \cap \dots \cap E_{n-1}) \end{aligned}$$

And we'd get

$$\begin{aligned} P(E_1 \cap \bar{E}_2 \cap E_3 \cap \dots \cap E_n) &= 1/11 \cdot 1/10 \cdot 1/9 \cdots 1/3 \cdot 1/2 \cdot 1/1 = 1/11! = |E_1 \cap E_2 \\ \cap E_3 \cap \dots \cap E_n|/N &= 1/N \implies N = 11! \end{aligned}$$

TREE DIAGRAMS



- Tree diagrams are a useful way of keeping track of the progress of compound experiments.
- They can also be seen to work using the multiplicative rule of probabilities.
- Let's analyse throwing 3 coins sequentially.

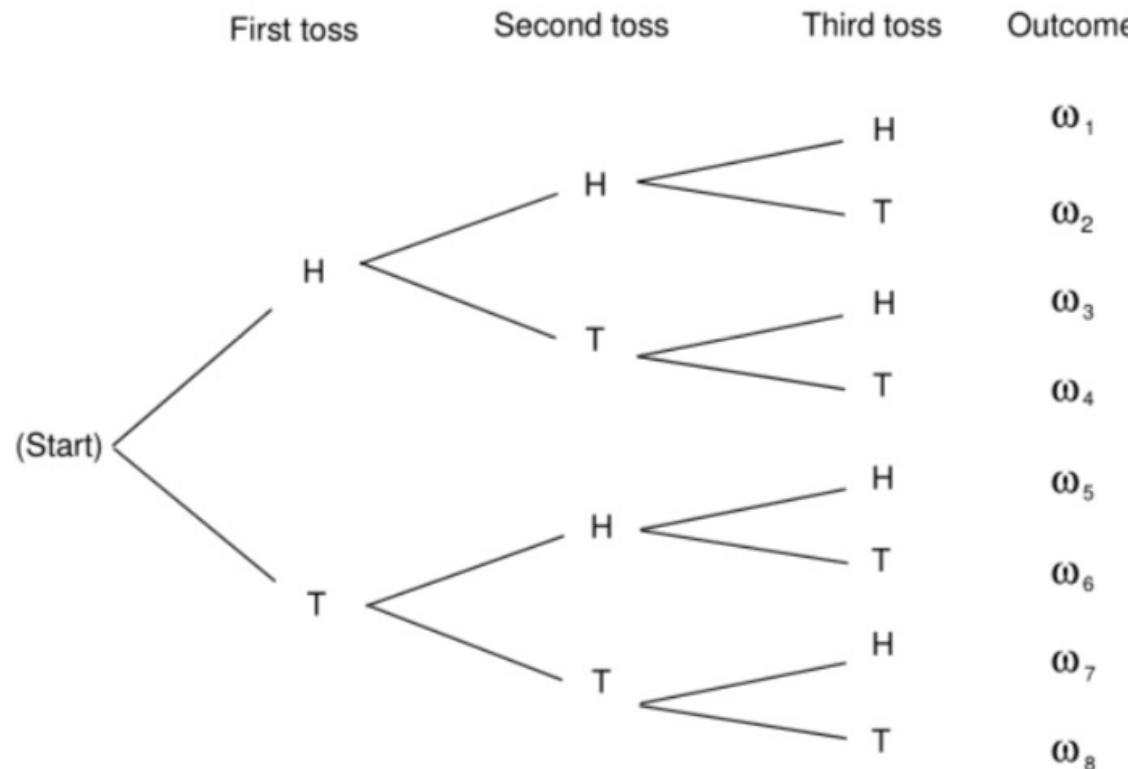
TREE DIAGRAMS

Tree diagrams are visual representations that help organize and illustrate the outcomes of compound experiments, especially when these experiments involve multiple stages or steps. They provide a clear and structured way of understanding the possible outcomes and the probabilities associated with each outcome.

In the context of probability theory, the multiplicative rule states that the probability of two independent events occurring together is the product of their individual probabilities. Tree diagrams are particularly useful for applying this rule because they help break down the compound experiment into its individual stages, making it easier to calculate probabilities at each step and then multiply them together to find the overall probability of a specific outcome.

For example, let's consider the experiment of tossing 3 coins sequentially.

TREE DIAGRAMS



Tree diagram for three tosses of a coin.

A tree diagram for this experiment would have three levels, each corresponding to a toss of a coin. At each level, there are two branches representing the two possible outcomes of a coin toss: heads (H) or tails (T). By following the branches of the tree diagram, we can systematically enumerate all possible sequences of coin toss outcomes.

TREE DIAGRAMS



Elaborating on how tree diagrams work with the multiplicative rule, at each stage of the experiment represented by a level in the tree diagram, we calculate the probability of each outcome (heads or tails) based on the probability of a single coin toss (which is $1/2$ for a fair coin).

Then, we multiply these probabilities along the branches to find the probability of each sequence of outcomes. Finally, we sum up the probabilities of all sequences that lead to the event of interest to determine its overall probability.

TREE DIAGRAMS

The overall probabilities to arrive at one outcome is obtained by multiplying the probabilities at each stage of the experiment.

This works because at each stage in the compound experiment, the result doesn't depend on the result further down the chain. So, to get

$$\begin{aligned}P(\omega_1) &= P(\{(H, H, H)\}) = P(\{H\}) \cdot P(\{(H, H)\} \mid \{H\}) \cdot P(\{(H, H, H)\} \mid \{(H, H)\}) \\&= P(\{\{H\}\}) \cdot P(\{\{H\}\}) \cdot P(\{\{H\}\}) = p^3\end{aligned}$$

because each probability $P(\{(H, H, H)\} \mid \{(H, H)\})$ only depends on the current coin flip, not anything else.

In other words, the individual parts of the compound experiment are independent.

TREE DIAGRAMS



EXAMPLE

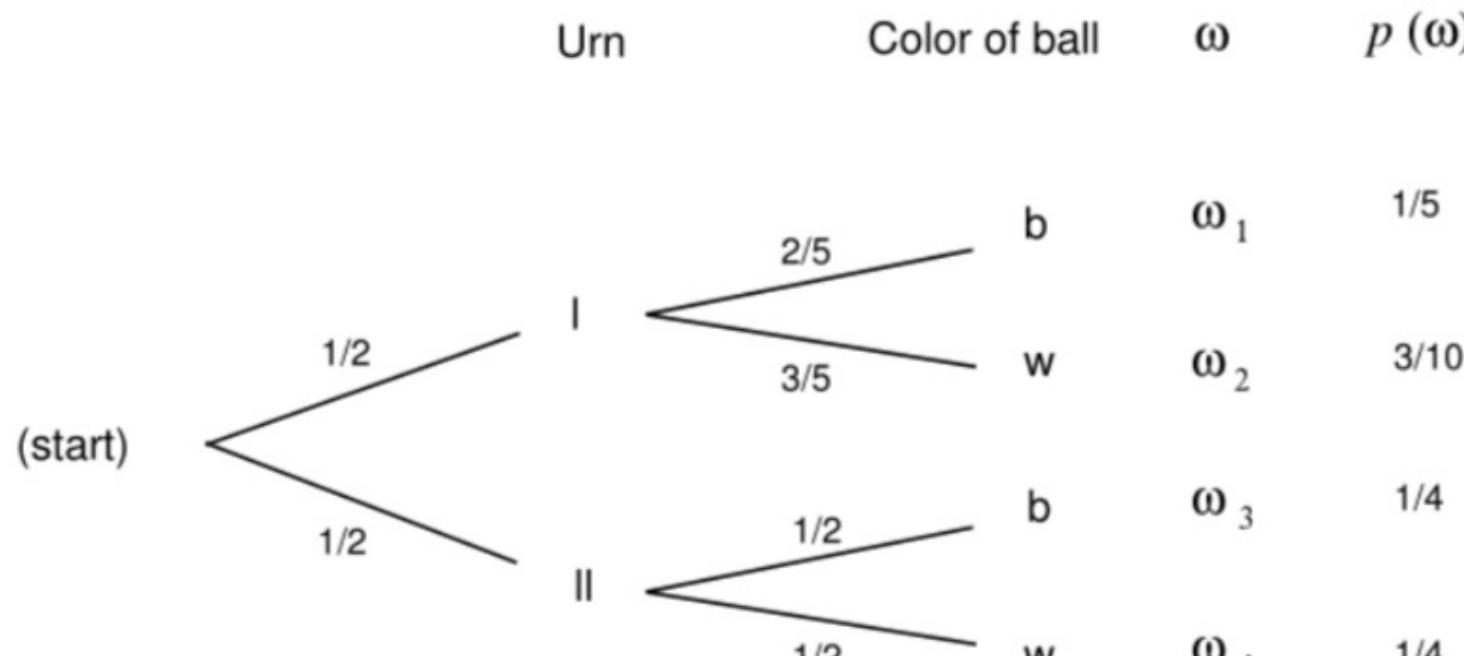
Consider two urns, labeled Urn I and Urn II.

- Urn I contains 2 black balls and 3 white balls.
- Urn II contains 1 black ball and 1 white ball.

Now, let's suppose that an urn is selected at random from the two available urns, and then a ball is chosen randomly from the selected urn.

TREE DIAGRAMS

We can present the sample space of this experiment as the paths through a tree as shown

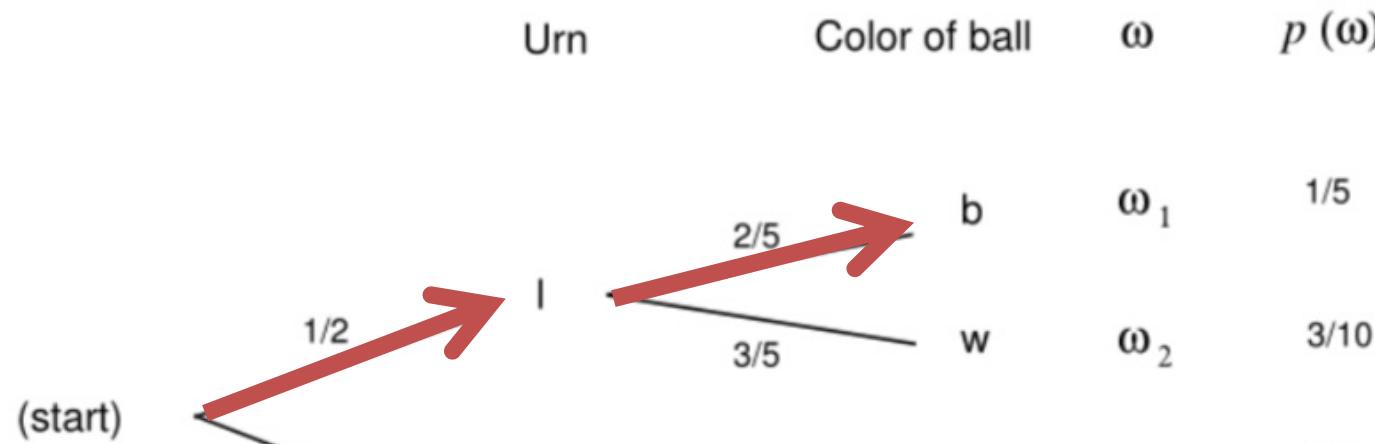


Tree diagram.

The probabilities assigned to the paths are also shown.

TREE DIAGRAMS

Let B = “a black ball is drawn,” and I = “urn I is chosen.”



Then the branch weight $2/5$, which is shown on one branch in the figure, can now be interpreted as the conditional probability

$$P(B | I).$$

If there were more branching's, we would have to use the multiplicative rule of probability.

EXAMPLE - BACK TO THE URNS AND BALLS

Suppose that we wanted to work out $P(I | B)$?

In words

"What is the probability that the ball came from urn I, given that the ball is black."

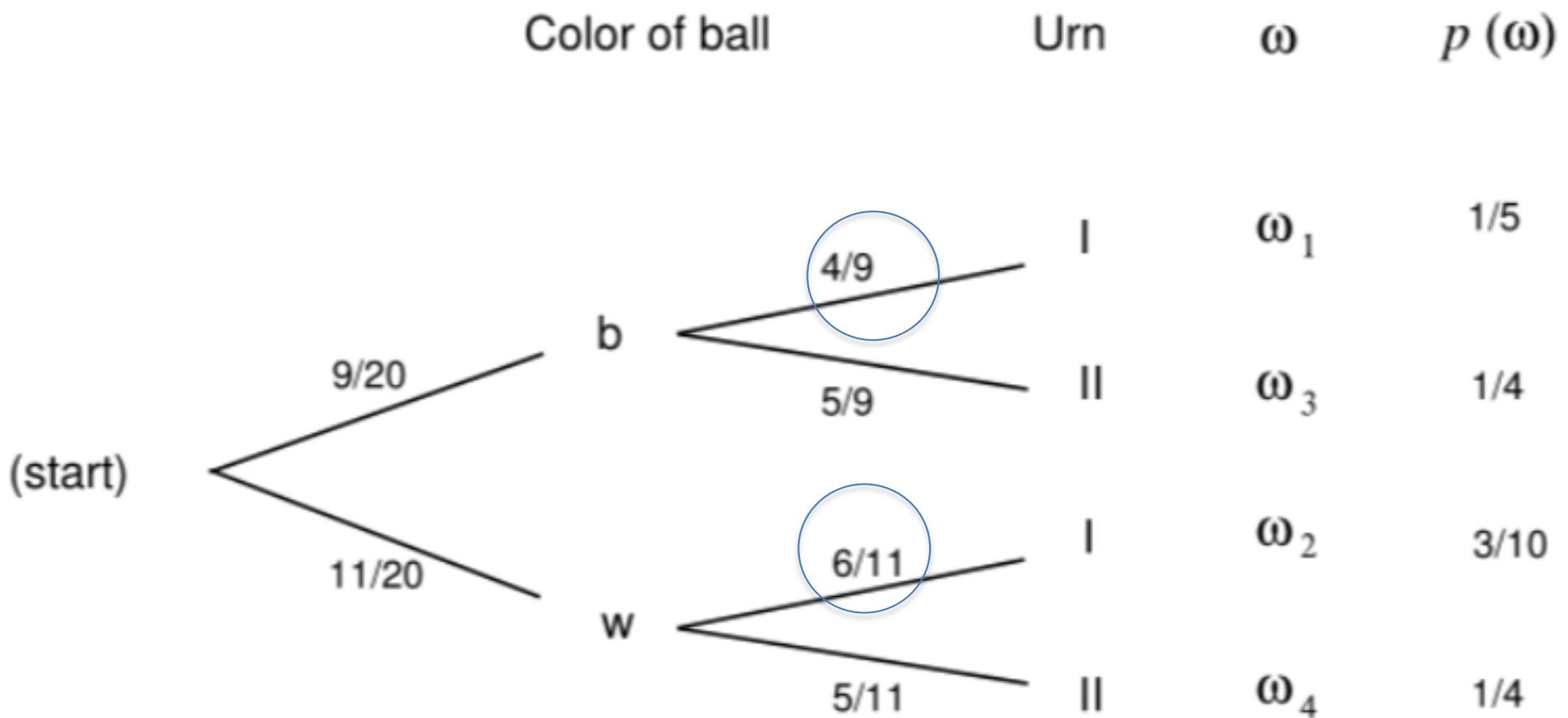
We can do this as follows

$$P(I | B) = \frac{P(I \cap B)}{P(B)} = \frac{P(I \cap B)}{P(B \cap I) + P(B \cap II)} = \frac{\frac{1}{5}}{\frac{1}{5} + \frac{1}{4}} = \frac{4}{9}$$

$$P(I|W) = \frac{P(I \cap W)}{P(W)} = \frac{P(I \cap W)}{P(W \cap I) + P(W \cap II)} = \frac{\frac{3}{10}}{\frac{3}{10} + \frac{1}{4}} = \frac{6}{11}$$

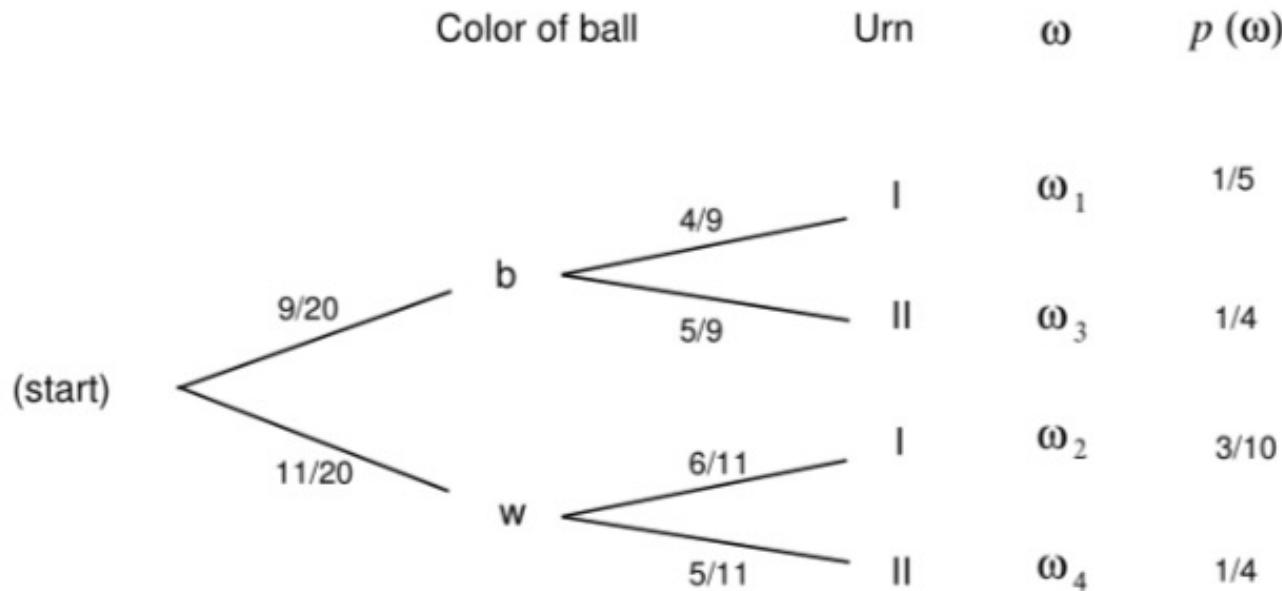
EXAMPLE - BACK TO THE URNS AND BALLS

We can repeat this for other conditional probabilities and use it to construct a reverse tree diagram



Reverse tree diagram.

EXAMPLE - BACK TO THE URNS AND BALLS



Reverse tree diagram.

These '*reversed*' conditional probabilities are often associated with the name **Bayes**.

We'll look at a formula that bears his name shortly that can be used to systematically calculate these reversed probabilities.

LAW OF TOTAL PROBABILITY

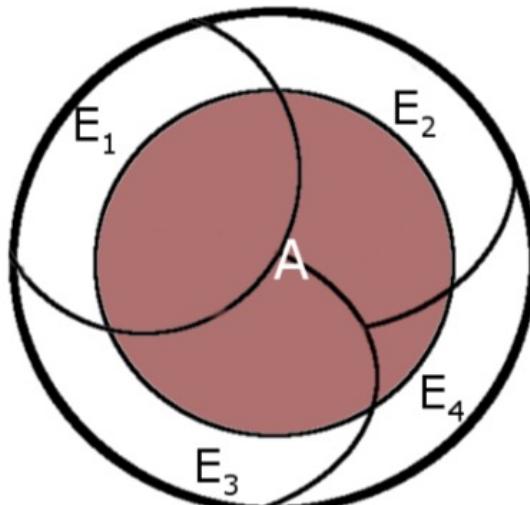
We can split S up into an exhaustive collection of mutually exclusive events E_i such that

$$\bigcup_{i=1}^n E_i = S$$

remembering that mutually exclusive means that

$$E_i \cap E_j = \emptyset, i \neq j.$$

In Venn diagram form this collection of events could look like this

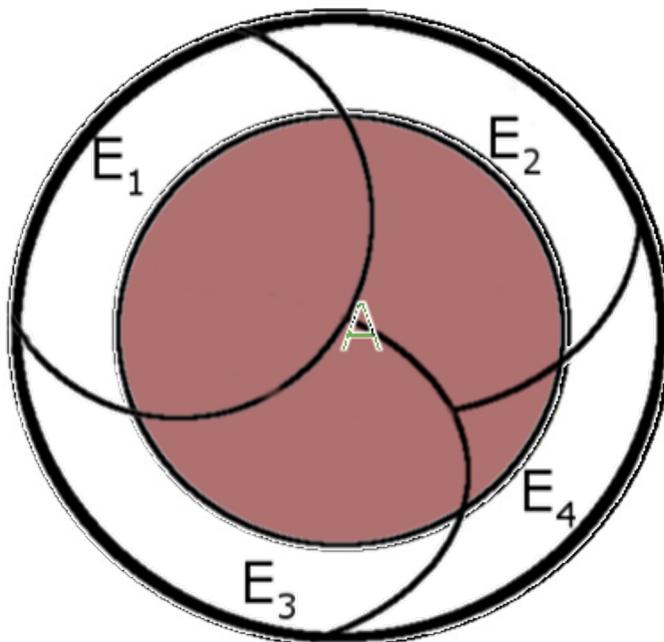


where $A \subset S$.

LAW OF TOTAL PROBABILITY

we can then use the E_i (a bit like basis vectors in linear algebra) to decompose an event A into its intersections with the E_i .

$$A = \bigcup_{i=1}^n A \cap E_i$$



LAW OF TOTAL PROBABILITY

From the conditional probability formula

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

Then, it follows that the general form of the law of total probability is given by the following expression

$$P(A) = \sum_{i=1}^n P(A \cap E_i) = \sum_{i=1}^n P(A | E_i)P(E_i)$$

BAYES' FORMULA

Using the law of total probability in the previous slides

$$P(A) = \sum_{i=1}^n P(A \cap E_i) = \sum_{i=1}^n P(A | E_i)P(E_i)$$



Rev. Thomas Bayes
(1701-1761)

$$P(E_i | A) = \frac{P(A \cap E_i)}{P(A)} = \frac{P(A | E_i)P(E_i)}{\sum_{j=1}^n P(A | E_j)P(E_j)}$$

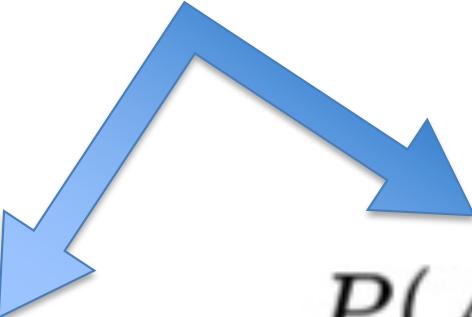
Bayes' solution to the problem of inverse probability was presented in "*An Essay Towards Solving a Problem in the Doctrine of Chances*," which was read to the Royal Society in 1763 following Bayes' death. This solution was further developed by Pierre-Simon Laplace, who first published the modern formulation in 1812. For a detailed account of the history of the Bayes formula, one can refer to the essay by E.T. Jaynes, "Bayesian Methods: General Background." (<https://bayes.wustl.edu/etj/articles/general.background.pdf>)



Pierre-Simon
Laplace
(1749-1827)

BAYES' FORMULA

If we look at the formula, we see that the two types of conditional probabilities have swapped events on the left and right hand sides.

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


BAYES' FORMULA



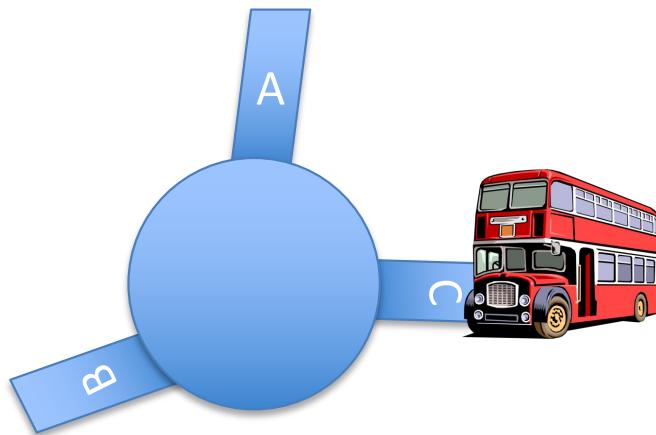
Bayes' Formula can be regarded as a method to revise our initial beliefs about a problem as new data becomes available. From this perspective:

- The E_i represents our prior beliefs or opinions about the likelihood of various events.
- These beliefs are then adjusted or updated based on the occurrence of event A.

This approach can also be applied to data-fitting tasks. In this context, Bayesian inference stands as the other primary approach, known for its subjective interpretation within probability theory.

BAYES' FORMULA: AN EXAMPLE

A town has three bus routes, A,B and C. *During rush hour there are twice as many buses on the A route as on B or C.*



Over a period of time, it has been observed that at a crossroads, where the routes converge, the buses run more than 5 minutes late $1/2$, $1/5$, $1/10$ of the time.

If an inspector at the crossroads finds that the first bus, he sees is more than five minutes late, what is the chance that it is a route B bus?

BAYES' FORMULA: AN EXAMPLE

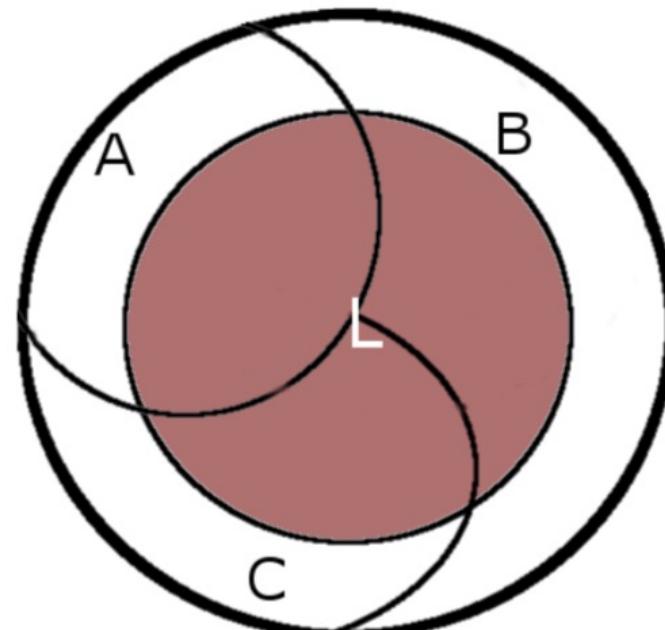
We don't have complete information about this problem (like all possible outcomes etc).

We can form an exhaustive partition of the experiment into the mutually exhaustive events A, B, C that the inspector sees a bus on that route first. And we have the event L that the bus is late.

We see that the problem looks like our Bayes' Formula situation: we know the probabilities

$$P(L|A), P(L|B), P(L|C),$$

but we want to know
 $P(B|L)$.



BAYES' FORMULA: AN EXAMPLE

We require $P(B | L)$. Using Bayes' formula and the law of the total probability, we get that

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

From the information we have ("during rush hour there are twice as many buses on the A route as on B or C") we can work out that

$$P(A) = 1/2 \text{ and } P(B) = P(C) = 1/4$$

BAYES' FORMULA: AN EXAMPLE



Also, we are given that :

Over some time, it has been observed that at a crossroads, where the routes converge, the buses run more than 5 minutes late 1/2 , 1/5 , 1/10 of the time.

That means:

$$P(L | A) = 1/2,$$

$$P(L | B) = 1/5,$$

$$P(L | C) = 1/10$$

Plugging these numbers in the formula gives $P(B | L) = 2/13$

BAYES' FORMULA: ANOTHER EXAMPLE



Bayes' rule applies not only to probabilistic scenarios but also to situations involving uncertainty regarding the value of a measured quantity.

For instance, let's consider an example from J.V. Stone's work, "*Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*," which aims to determine the probability of correctly understanding the meaning of a spoken conversation.

Before delving into this example, it might be helpful to provide some context, such as the famous gag from The Two Ronnies...



BAYES' FORMULA: ANOTHER EXAMPLE



In any spoken language, regional variations in pronunciation can lead to humorous misunderstandings, as famously depicted in this sketch of *The Two Ronnies*.

Now, let's explore how Bayes' Rule can be employed to calculate the probability of providing an incorrect answer in such situations.

BAYES' FORMULA: ANOTHER EXAMPLE



BAYES' FORMULA: ANOTHER EXAMPLE

Likelihood

$$P(\text{acoustic data} \mid \text{four candles}) = 0.6$$

$$P(\text{acoustic data} \mid \text{fork handles}) = 0.7$$



This give the probabilities of acoustic data given the two possible phrases.
In this case it is almost identical.

Each likelihood answer to the question:

What is the probability of the observed acoustic data given that each of two possible phrases was spoken?

Posterior probability

$$P(\text{four candles} \mid \text{data})$$

$$P(\text{fork handles} \mid \text{data})$$



This give the probability of each phrase given the acoustic data

BAYES' FORMULA: ANOTHER EXAMPLE

Prior probability

Let's suppose that the assistant has been asked for four candle a total of 90 times
In the past, whereas he has been asked for fork handles only 10 times.

Therefore, the assistant estimates that the probability for a next client asking for one of the two items are

$$P(\text{four candles}) = \frac{90}{100} = 0.9$$

$$P(\text{fork handles}) = \frac{10}{100} = 0.1$$

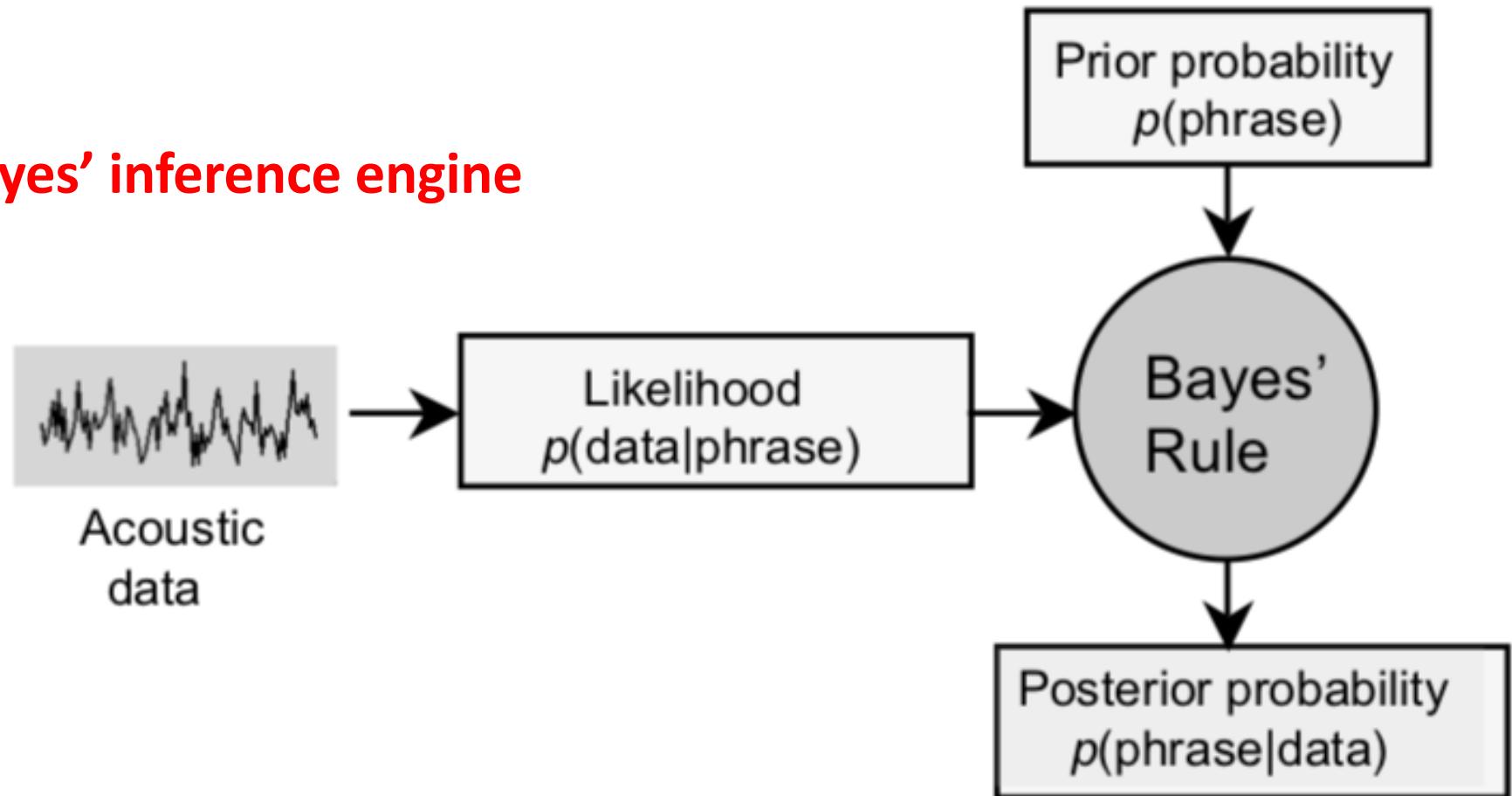


This give the prior knowledge of the shop assistant
Based on his previous experience of what customers say

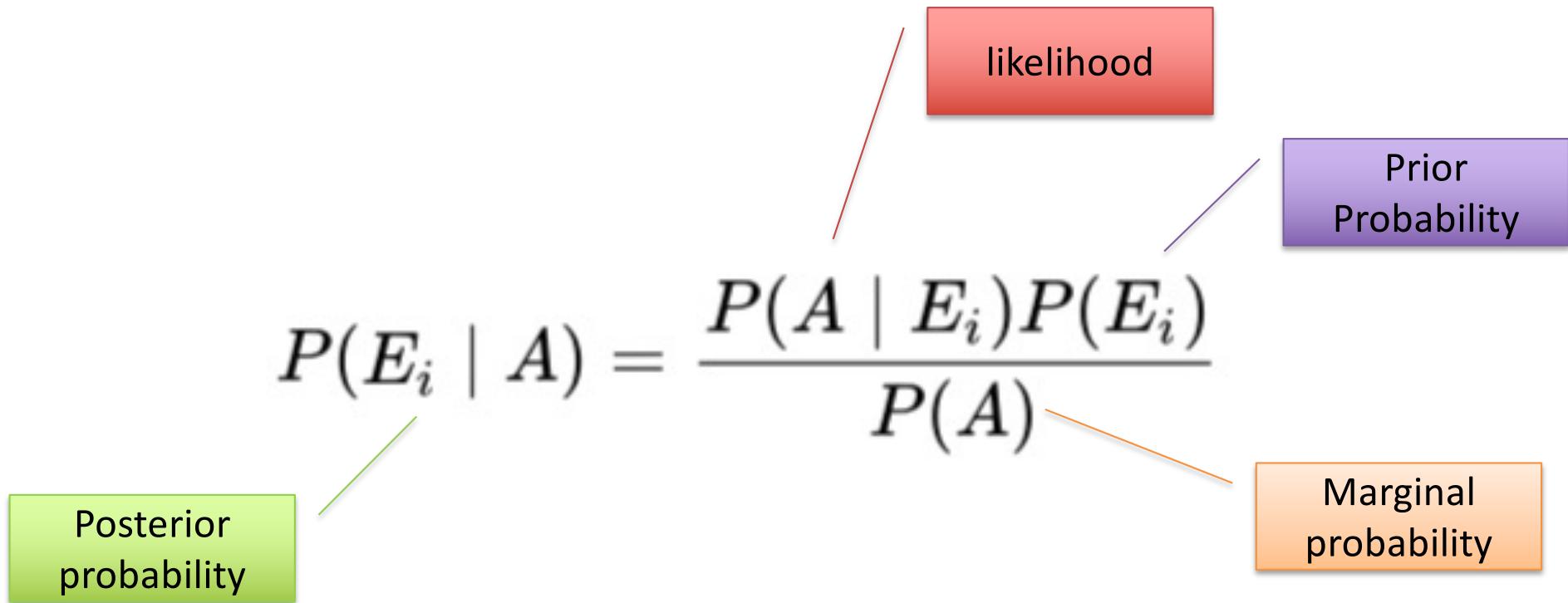
BAYES' FORMULA: ANOTHER EXAMPLE

INFERENCE : *A conclusion reached on the basis of evidence and reasoning.*

Bayes' inference engine



BAYES' FORMULA: ANOTHER EXAMPLE

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


Posterior probability

likelihood

Prior Probability

Marginal probability

$$p(\text{four candles}|\text{data}) = \frac{p(\text{data}|\text{four candles})p(\text{four candles})}{p(\text{data})}$$

$$p(\text{fork handles}|\text{data}) = \frac{p(\text{data}|\text{fork handles})p(\text{fork handles})}{p(\text{data})},$$

BAYES' FORMULA: ANOTHER EXAMPLE

Using the calculate probability we can estimate the posterior probability

$$\begin{aligned} p(\text{four candles}|\text{data}) &= p(\text{data}|\text{four candles})p(\text{four candles})/p(\text{data}) \\ &= 0.6 \times 0.9/0.61 = 0.885, \end{aligned}$$

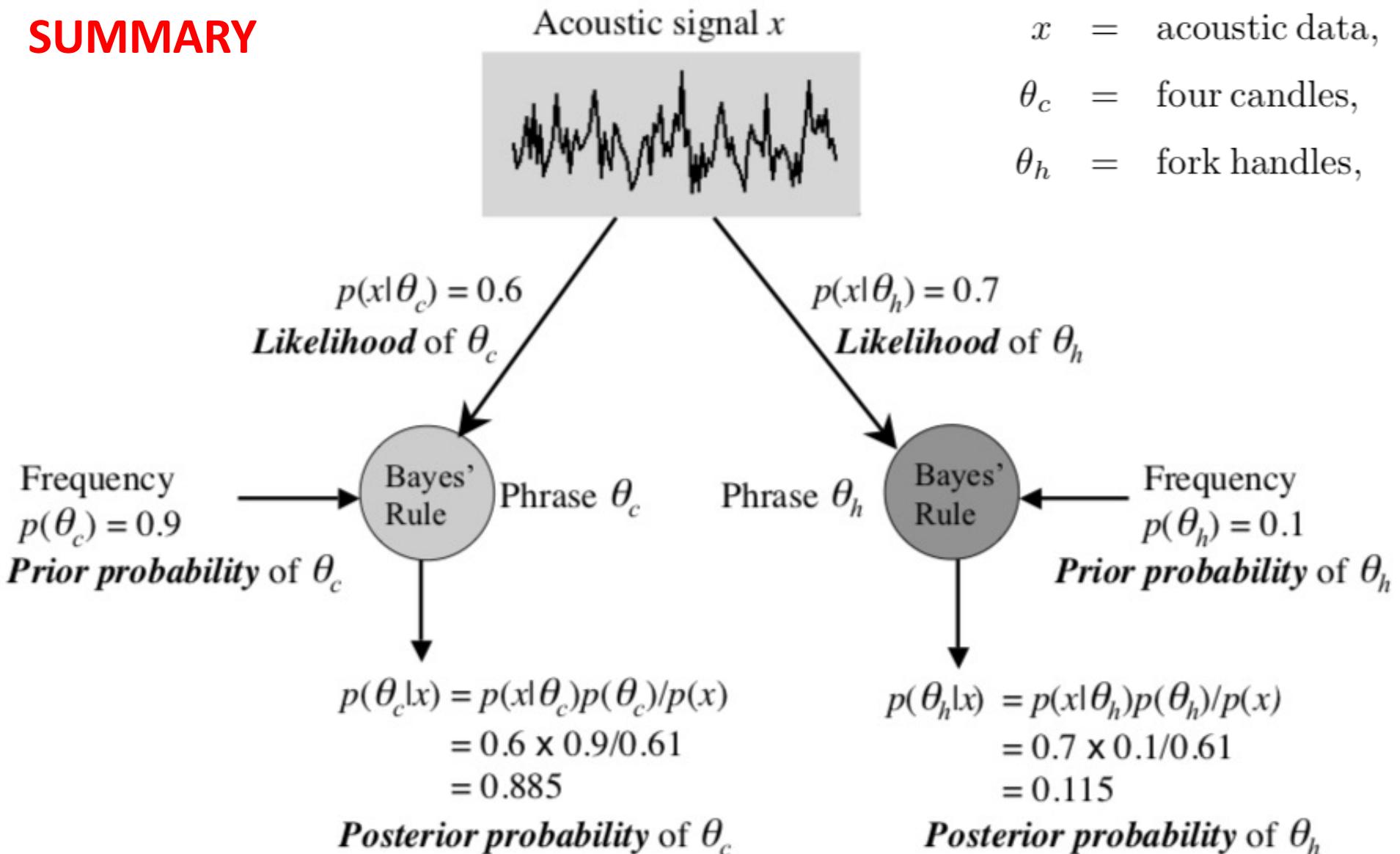
$$\begin{aligned} p(\text{fork handles}|\text{data}) &= p(\text{data}|\text{fork handles})p(\text{fork handles})/p(\text{data}) \\ &= 0.7 \times 0.1/0.61 = 0.115. \end{aligned}$$

$$p(\theta_c|x) = p(x|\theta_c)p(\theta_c)/p(x) = 0.885 \quad x = \text{acoustic data},$$

$$p(\theta_h|x) = p(x|\theta_h)p(\theta_h)/p(x) = 0.115. \quad \theta_c = \text{four candles},$$
$$\theta_h = \text{fork handles},$$

BAYES' FORMULA: ANOTHER EXAMPLE

SUMMARY



CONDITIONAL PROBABILITIES ARE PROBABILITIES

It is important to note that conditional probabilities obey all the axioms we established to call $P(E)$ a probability.

For the statement above to be true our conditional probabilities must satisfy the axioms of probabilities:

- $0 \leq P(E | F) \leq 1$
- $P(S) = 1$
- if $E_i, i = 1 \dots n$ are mutually exclusive events then

$$P\left(\bigcup_{i=1}^n E_i | F\right) = \sum_{i=1}^n P(E_i | F)$$

We are not going to prove these results here, but you can find a proof in textbook.

SUMMARY OF CONDITIONAL PROBABILITIES



- **Definition of conditional probabilities**

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)}$$

- **Multiplicative rule**

$$\begin{aligned} P(E_1 E_2 E_3 \dots E_n) &= P(E_1) \frac{P(E_1 \cap E_2)}{P(E_1)} \frac{P(E_1 \cap E_2 \cap E_3)}{P(E_1 \cap P(E_2))} \\ &\quad \dots \frac{P(E_1 \cap E_2 \dots \cap E_n)}{P(E_1 \cap E_2 \dots \cap E_{n-1})} \end{aligned}$$

SUMMARY OF CONDITIONAL PROBABILITIES



- Law of total probability

$$P(A) = \sum_{i=1}^n P(A \cap E_i) = \sum_{i=1}^n P(A | E_i)P(E_i)$$

- Bayes' Formula

$$P(E_i | A) = \frac{P(A \cap E_i)}{P(A)} = \frac{P(A | E_i)P(E_i)}{\sum_{j=1}^n P(A | E_j)P(E_j)}$$

INDEPENDENT EVENTS



In the previous lesson, we discussed examples of experiments involving independent events.

Now that we have introduced conditional probability, let's delve deeper into this concept.

INDEPENDENT EVENTS



Two experiments are considered independent if the outcome of one experiment has no impact on the possible outcomes of the other.

Similarly, two events (E, F) are termed independent if the occurrence of one event does not influence the probability of the other event happening.

INDEPENDENT EVENTS



Mathematically, this independence can be expressed through conditional probabilities:

- For event E, the probability of E occurring is equal to the conditional probability of E given that event F has occurred, and also equal to the conditional probability of E given that event F has not occurred. This is represented as: $P(E) = P(E|F) = P(E|\neg F)$
- Similarly, for event F, the probability of F occurring is equal to the conditional probability of F given that event E has occurred, and also equal to the conditional probability of F given that event E has not occurred. This is represented as: $P(F) = P(F|E) = P(F|\neg E)$

INDEPENDENT EVENTS



Another way to interpret this is that the probability of both events E and F occurring (denoted as $P(E \cap F)$) is simply the product of the probabilities of each event occurring:

$$P(E \cap F) = P(E) \cdot P(F)$$

In simpler terms, if events E and F are independent, then the likelihood of both events happening simultaneously is just the product of the likelihood of each event happening independently.

INDEPENDENT EVENTS



Theorem: Two events (E, F) are independent if and only if their intersection ($E \cap F$) equals the product of their probabilities:

$$P(E \cap F) = P(E)P(F)$$

INDEPENDENT EVENTS



However, when dealing with more than two events, the notion of independence becomes more restrictive.

Theorem: *a set of events $E_1, E_2, E_3, \dots, E_n$ are considered mutually independent if for every subset if for every subset $E'_1, E'_2, E'_3, \dots, E'_r, r \leq n$*

the probability of the intersection of all events in the subset equals the product of the probabilities of each individual event in the subset:

$$P(E'_1 \cap E'_2 \cap E'_3 \cap \dots \cap E'_r) = P(E'_1) \cdot P(E'_2) \cdot P(E'_3) \cdots P(E'_r)$$

INDEPENDENT EVENTS



The concept of independent processes holds significant importance as we progress further.

We frequently rely on the notion that repetitions of experiments constitute independent processes.

It's essential to note that this assumption aligns closely with the frequentist definition of probability.

INDEPENDENT EVENTS: EXAMPLE

Consider the compound experiment of throwing two fair coins.
The sample space is

$$S = \{(H, H), (H, T), (T, H), (T, T)\}.$$

Define two events

$$A = 'the first coin is a head' = \{(H, H), (H, T)\}$$

$$B = 'the second coin is a tail' = \{(H, T), (T, T)\}$$

$$\text{And } P(A) = |A|/|S| = 2/4 = \frac{1}{2} \text{ and } P(B) = |B|/|S| = 2/4 = \frac{1}{2}$$

$$\text{Now } P(A \cap B) = P(H, T) = |A \cap B|/|S| = 1/4 = P(A) * P(B).$$

so A and B are independent.

INDEPENDENT EVENTS: EXAMPLE



But consider C = 'both coins are heads' = $\{(H, H)\}$

$$P(C) = |C|/|S| = 1/4$$

now

$$P(B \cap C) = P(\emptyset) = |B \cap C|/|S| = 0/4 \neq P(A) * P(C) = 1/2 * 1/4,$$

so B and C are not independent. A is also not independent of C .

In general, it is possible for all pairs of events to be independent, but the complete set of events not to be.



QUESTIONS





UNIVERSITY OF
LINCOLN

SCHOOL OF MATHEMATICS
AND PHYSICS

MTH1005

PROBABILITY AND STATISTICS

Semester B
Lecture 4
(20/2/2024)

Danilo Roccatano

Office:

INB 3323

Email:

droccatano@lincoln.ac.uk

SUMMARY OF CONDITIONAL PROBABILITIES

- **Definition of conditional probabilities for two events (E,F) in S**

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

- **Multiplicative rule for a set of events (E₁... E_n) in S**

$$P(E_1 E_2 E_3 \dots E_n) = P(E_1) \frac{P(E_1 \cap E_2)}{P(E_1)} \frac{P(E_1 \cap E_2 \cap E_3)}{P(E_1 \cap P(E_2))} \dots \frac{P(E_1 \cap E_2 \dots \cap E_n)}{P(E_1 \cap E_2 \dots \cap E_{n-1})}$$

SUMMARY OF CONDITIONAL PROBABILITIES



- Law of total probability for a set of events ($E_1 \dots E_n$) in S

$$P(A) = \sum_{i=1}^n P(A \cap E_i) = \sum_{i=1}^n P(A | E_i)P(E_i)$$

- Bayes' Formula for a set of events E_i

$$P(E_i | A) = \frac{P(A \cap E_i)}{P(A)} = \frac{P(A | E_i)P(E_i)}{\sum_{j=1}^n P(A | E_j)P(E_j)}$$

INDEPENDENT EVENTS



Definition: Two experiments are independent if the result of one cannot in any way affect the possible results of the other.

Definition: Two events (E, F) are independent if the probability that one of them occurs is in no way influenced by whether or not the other has occurred.

$$\begin{aligned} P(E) &= P(E \mid F) = P(E \mid \bar{F}), \\ P(F) &= P(F \mid E) = P(F \mid \bar{E}). \end{aligned}$$

put in a different way this means that

$$P(E \cap F) = P(E)P(F)$$

the probability of E and F occurring is just the product of the probability of E occurring and the probability of F occurring.

INDEPENDENT EVENTS

Theorem: Two events (E, F) are independent if and only if

$$P(E \cap F) = P(E)P(F)$$

For more than two events things become a bit more restrictive.

Theorem: The events $E_1, E_2, E_3, \dots, E_n$ are said to be mutually independent if for every subset

$$E'_1, E'_2, E'_3, \dots, E'_r, r \leq n$$

$$P(E'_1 \cap E'_2 \cap E'_3 \cap \dots \cap E'_r) = P(E'_1) \cdot P(E'_2) \cdot P(E'_3) \cdots P(E'_r)$$

INDEPENDENT EVENTS



The idea of independent processes will be extremely important as we move forward.

We will use the idea that repetitions of experiments constitute independent processes very often.

Note that this is more or less an assumption of the frequentist definition of probability.

INDEPENDENT EVENTS: EXAMPLE

Consider the compound experiment of throwing two fair coins.
The sample space is

$$S = \{(H, H), (H, T), (T, H), (T, T)\}.$$

Define two events

$$A = 'the first coin is a head' = \{(H, H), (H, T)\}$$

$$B = 'the second coin is a tail' = \{(H, T), (T, T)\}$$

$$P(A) = |A|/|S| = 2/4 = \frac{1}{2} \text{ and}$$

$$P(B) = |B|/|S| = 2/4 = \frac{1}{2}$$

$$\text{Now } P(A \cap B) = P(H, T) = |A \cap B|/|S| = 1/4 = P(A) * P(B).$$

so A and B are independent.

INDEPENDENT EVENTS: EXAMPLE



But consider $C = \text{'both coins are heads'} = \{(H, H)\}$

$$P(C) = |C|/|S| = 1/4$$

now

$$P(B \cap C) = P(\emptyset) = |B \cap C|/|S| = 0/4 \neq P(B) * P(C) = (1/2)(1/4),$$

$$P(A \cap C) = P(\emptyset) = |A \cap C|/|S| = 0/4 \neq P(A) * P(C) = (1/2)(1/4),$$

so B and C are not independent. A is also not independent of C .

In general, it is possible for all pairs of events to be independent, but the complete set of events not to be.

TABULAR PRESENTATION OF CONDITIONAL PROBABILITIES



It can sometimes be handy to view conditional probabilities using a tabular representation of relative frequencies - this can also be how real data arrives to us.

TABULAR PRESENTATION OF CONDITIONAL PROBABILITIES



THE BUS EXAMPLE

A town has three bus routes, A, B, and C. During rush hour there are twice as many buses on the A route as on B or C.

Over time, it has been observed that at a crossroads, where the routes converge, the buses run more than 5 minutes late $\frac{1}{2}$, $\frac{1}{5}$, $\frac{1}{10}$ of the time.

If an inspector at the crossroads finds that the first bus, he sees is more than five minutes late, what is the chance that it is a route B bus?

TABULAR PRESENTATION OF CONDITIONAL PROBABILITIES

If we go back to the bus problem we did the last lecture, but instead we consider what we'd expect if **1000** buses in total ran through the town, we'd end up with something like

	A	B	C	total
Late	250	50	25	325
Not late	250	200	225	675
Total	500	250	250	1000

TABULAR PRESENTATION OF CONDITIONAL PROBABILITIES

Let L = 'a bus is late'

	A	B	C	total
Late	250	50	25	325
Not late	250	200	225	675
Total	500	250	250	1000

The conditional probabilities can easily be read off, for instance

$$P(B | L) = P(B \cap L)/P(L) = 50/325 = 2/13$$

This is of course exactly equivalent to our pen and paper solution.

TABULAR PRESENTATION OF CONDITIONAL PROBABILITIES

	A	B	C	total
Late	250	50	25	325
Not late	250	200	225	675
Total	500	250	250	1000

Marginal
Probabilities

This is because of the law of total probability.

$$P(A) = \sum_{i=1}^n P(A \cap E_i) = \sum_{i=1}^n P(A | E_i)P(E_i)$$

TABULAR PRESENTATION OF CONDITIONAL PROBABILITIES

	A	B	C	total	
Late	250	50	25	325	
Not late	250	200	225	675	
Total	500	250	250	1000	

Another example

$$P(L) = P(L \cap A) + P(L \cap B) + P(L \cap C) = 250 + 50 + 25 = 325$$

or **vertically** we have

$$P(A) = P(A \cap L) + P(A \cap \bar{L})$$

note that that last relationship is a useful and general one.
It is a special case of the law of total probability.

TABULAR BAYES' THEOREM: yet another TABULAR EXAMPLE



A doctor is trying to decide if a patient has one of three diseases d_1 , d_2 , or d_3 .

- **Two tests** are to be carried out, each of which results in a positive (+) or a negative (-) outcome.
- There are four possible test patterns **++**, **+−**, **−+**, and **−−**.
- National records have indicated that, for 10,000 people having one of these three diseases, the distribution of diseases and test results are as in the table in the next slide.

TABULAR BAYES' THEOREM: yet another TABULAR EXAMPLE

Disease	number	++	+ -	- +	--
d1	3215	2110	301	704	100
d2	2125	396	132	1187	410
d3	4660	510	3568	73	509
Total	10000				

We can use this data to estimate $P(d_1)$, $P(d_2)$, $P(d_3)$ - these are called prior probabilities, and the conditional probabilities like

$$P(+ - | d_1) = 301/3215 = 0.094$$

TABULAR BAYES' THEOREM: yet another TABULAR EXAMPLE

What the doctor wants though is

the probability a patient has disease d_i , given the results of the tests.

These are the ***Bayes' or inverse, or posterior probabilities.***

We can compute them using Bayes' formula and we'll get results like

- Bayes' Formula

$$P(E_i | A) = \frac{P(A \cap E_i)}{P(A)} = \frac{P(A | E_i)P(E_i)}{\sum_{j=1}^n P(A | E_j)P(E_j)}$$

Ex.

$$P(d_1 |++) = P(++)|d_1)P(d_1)/P(++) = 0.700 .$$

	d_1	d_2	d_3
$++$.700	.131	.169
$+ -$.075	.033	.892
$- +$.358	.604	.038
$--$.098	.403	.499

TABULAR BAYES' THEOREM: A TABULAR EXAMPLE

The values in the table are $P(d_i, ++)$ etc. Judicious use of these posterior probabilities can be used to inform decision:

	d_1	d_2	d_3
$++$.700	.131	.169
$+ -$.075	.033	.892
$- +$.358	.604	.038
$--$.098	.403	.499

If the test result came back $++$ then the posterior probability $P(d_1, ++) = 0.7$ and we'd suspect that d_1 was the culprit.

UPDATING INFORMATION SEQUENTIALLY



- In the last example, we updated the likelihood of a patient having a particular disease based on extra information in the form of the test results.
- The example should also give a hint about how we can update our ideas about a system as new information comes in.
- this is why the original probabilities are called priors, and the reversed conditional probabilities are called posterior probabilities.
- *We'll use this type of method in the computing lab later as a simple form of machine learning.*



RANDOM VARIABLES

RANDOM VARIABLES



In the next sessions we will define, discuss and begin to use random variables.

Understand the elements of a random variable.
Be able to define and use

- Discrete vs continuous variables
- Range and domain of the variables
- Probability (Mass/Density) Function
- Cumulative Distribution Function Improper integrals

RANDOM VARIABLES



Random variables are a tool that let us work with more complicate situations using still using a lot of tools that we know from algebra or calculus.

Unlike a normal variable, it doesn't make sense to say ' $x = \text{some value}$ '.

For example:

- How many miles to a gallon does my car do?
- How quickly do I run a mile ?
- How long will it take for a website to load?
- What will be the average daily temperature in two days?

RANDOM VARIABLES



Random variables inherently can take on a variety of values, and only discussing a statistic distribution of their values makes sense. Every time we tried something (took a measurement) we would get a slightly different answer.

Most idealisations of real-life experiments would correspond to random variables!

We'll see

- what random variables are
- how they can link abstract probability distributions to real-life experiments

STRING LOGIC



The random variable is a series of 'mappings', from events to probabilities to numerical values.

We

- split the sample space into events,
- we assign each event a probability,
- we assign each event a unique numerical value.

the random variable is the whole of that package.

STRING LOGIC



Lets consider rolling a normal fair six sided die.

We want to quantify the number of spots on the topmost face of die after a single roll.

If we had an infinite amount of information we could determine which face will come up before the roll was made. But as we don't we can only work out the probabilities of the different faces coming up.

With those different faces coming up, and their probabilities we associate numbers.

STRING LOGIC

These steps can be done in any order, but often it makes sense to assign the values we want to the random variable.

We'll call our random variable Z.

We'll mix up some mathematical notation and common sense

$$Z = \begin{cases} 1 & \text{if 'the die lands with 1 spot on the top face'} \\ 2 & \text{if 'the die lands with 2 spots on the top face'} \\ 3 & \text{if 'the die lands with 3 spots on the top face'} \\ 4 & \text{if 'the die lands with 4 spots on the top face'} \\ 5 & \text{if 'the die lands with 5 spots on the top face'} \\ 6 & \text{if 'the die lands with 6 spots on the top face'} \end{cases}$$

the statements if 'the die lands with 1 spot on the top face' hopefully sound a lot like events to you.

STRING LOGIC

$$S = \{1, 2, 3, 4, 5, 6\}$$

where $\{1\}$ is the outcome that the die lands with 1 spot on the top face, $\{2\}$ is the outcome that the die lands with 2 spots up etc.

So looking again at Z

$$Z = \begin{cases} 1 & \text{if 'the die lands with 1 spot on the top face'} \\ 2 & \text{if 'the die lands with 2 spots on the top face'} \\ 3 & \text{if 'the die lands with 3 spots on the top face'} \\ 4 & \text{if 'the die lands with 4 spots on the top face'} \\ 5 & \text{if 'the die lands with 5 spots on the top face'} \\ 6 & \text{if 'the die lands with 6 spots on the top face'} \end{cases}$$

we could replace the text 'the die lands with 1 spot on the top face' with $\{1\}$, the event that the die lands with 1 spot upwards, and get

STRING LOGIC

$$Z = \begin{cases} 1 & \text{if } \{1\} \\ 2 & \text{if } \{2\} \\ 3 & \text{if } \{3\} \\ 4 & \text{if } \{4\} \\ 5 & \text{if } \{5\} \\ 6 & \text{if } \{6\} \end{cases}$$

STRING LOGIC



Now what is the probability that we get those different values of Z ?

We get $Z = 1$ if the event $\{1\}$, or 'the die lands with 1 spot on the top face' occurs.

You should know how to associate a probability to an event occurring.

In this case where all the outcomes $\{1, 2, 3, 4, 5, 6\}$ can be argued to be equally likely, the probability of any one of them occurring must be $1/|S|$, one over the number of elements in the sample space, in this case $1/6$.

So the probability that $Z= 1$, which is that $\{1\}$ occurred is $1/6$

STRING LOGIC

we'll define another function, the probability (mass) function of Z .

$$p_Z(z) = P(Z = z) = \begin{cases} 1/6 & \text{if } \{1\} \\ 1/6 & \text{if } \{2\} \\ 1/6 & \text{if } \{3\} \\ 1/6 & \text{if } \{4\} \\ 1/6 & \text{if } \{5\} \\ 1/6 & \text{if } \{6\} \end{cases}$$

and our random variable Z is fully defined.

EXAMPLE OF 6-sided DIE

we have our sample space

$$S = \{1, 2, 3, 4, 5, 6\}$$

the values of our random variable, and their associated probabilities

$$Z = \begin{cases} 1 & \text{if } \{1\} \\ 2 & \text{if } \{2\} \\ 3 & \text{if } \{3\} \\ 4 & \text{if } \{4\} \\ 5 & \text{if } \{5\} \\ 6 & \text{if } \{6\} \end{cases}, p_Z(z) = P(Z = z) = \begin{cases} 1/6 & \text{if } \{1\} \\ 1/6 & \text{if } \{2\} \\ 1/6 & \text{if } \{3\} \\ 1/6 & \text{if } \{4\} \\ 1/6 & \text{if } \{5\} \\ 1/6 & \text{if } \{6\} \end{cases}$$

we must have that each possible value of the random variable is unambiguously associated with an event and a probability.
It should be clear that this is the case above.

DEFINITION OF DISCRETE RANDOM VARIABLE

Definition Let X be a discrete random variable that can take on only the values $x_1, x_2, x_3, \dots, x_n$ with respective probabilities $p_X(x_1), p_X(x_2), p_X(x_3), \dots, p(x_n)$. Then, if

$$\sum_{x \in X} p_X(x) = 1,$$

X is a **discrete random variable**.

The function $p(x) = P\{X = x\}$ is called the probability (mass) function of the variable X .

NOTE: Other notations are sometimes used for the probability (mass) function, $m(i)$ or p_i .

It is also conventional that random variables are given a capital letter, and normally are chosen from the end of the alphabet).

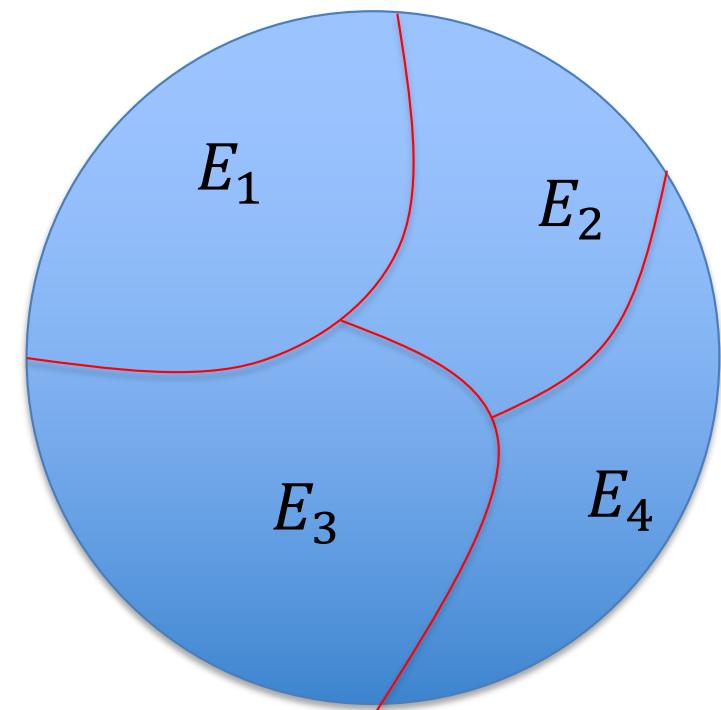
DISCRETE RANDOM VARIABLES

Definition: Consider an experiment, with outcome set S , split into n mutually exclusive and exhaustive events $E_1, E_2, E_3, \dots, E_n$. A variable, X say, which can assume exactly n numerical values each of which corresponds to one and only one of the given events is called a random variable.

Schematically the mutually exclusive and exhaustive events look like

Event types:

- Exclusive: no overlap.
- Exhaustive: all of S is covered by them.



DISCRETE RANDOM VARIABLES



Here our outcome set is split into 4 mutually exclusive events and exhaustive. So to associate a random variable (call it X) with this sample space we could have something like

$X = 1$, corresponding to event E_1

$X = 2$, corresponding to event E_2

$X = 3$, corresponding to event E_3

$X = 4$, corresponding to event E_4

and we know how to calculate the probability of events.

STRING LOGIC



The random variable is a series of 'mappings', from events to probabilities to numerical values.

We

- split the sample space into events,
- we assign each event a probability,
- we assign each event a unique numerical value.

the random variable is the whole of that package.

STRING LOGIC



Check list

- is the sample space well defined?
- are the events mutually exclusive?
- do the events cover all the sample space?
- are the probabilities of the events defined
- are values of the variable clearly assigned one-to-one to the possible events?

EXAMPLE

Define a valid random variable to describe the number of girls in families with 2 children, assuming that the likelihoods of boys or girls is equal and independent.

First, we define our sample space

$$S = \{(B, B), (B, G), (G, B), (G, G)\}$$

We define our variable G using the number of girls in the family:

$$G = \begin{cases} 0 & \text{if } \{(B, B)\} \\ 1 & \text{if } \{(B, G), (G, B)\} \\ 2 & \text{if } \{(G, G)\} \end{cases}$$

EXAMPLE

and associate our probability (mass) function

$$p_G(x) = \begin{cases} p_G(0) = P(G = 0) = P(\{(B, B)\}) = \frac{1}{4} \\ p_G(1) = P(G = 1) = P(\{(B, G), (G, B)\}) = \frac{1}{2} \\ p_G(2) = P(G = 2) = P(\{(G, G)\}) = \frac{1}{4} \end{cases}$$

We check that

$$\sum_{g \in G} p_G(g) = 1.$$

Here g indicates the possible values of G , so

$$\sum_{g \in G} p_G(g) = p_G(0) + p_G(1) + p_G(2) = 1/4 + 1/2 + 1/4 = 1$$

EXAMPLE 2



Construct a random variable that counts the total number of spots upwards when two fair 6-sided dice are thrown. Our sample space can be the 36 elements of

$$S = \{(i, j) : i, j = 1, 2, 3, 4, 5, 6\}$$

where i is the number of spots on the first die, and j the second die.

EXAMPLE 2

How we need to think of a set of mutually exclusive and exhaustive events that would correspond to the sum of the spots. We could write this in set builder notation

$$E_k = \{ (i, j) : i + j = k; i, j = 1, 2, 3, 4, 5, 6 \}$$

and we can write these out in full

$$E_2 = \{(1, 1)\}$$

$$E_3 = \{(1, 2), (2, 1)\}$$

$$E_4 = \{(1, 3), (2, 2), (3, 1)\}$$

$$E_5 = \{(1, 4), (2, 3), (3, 2), (4, 1)\}$$

$$E_6 = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$$

$$E_7 = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$$

$$E_8 = \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$$

$$E_9 = \{(3, 6), (4, 5), (5, 4), (6, 3)\}$$

$$E_{10} = \{(4, 6), (5, 5), (6, 4)\}$$

$$E_{11} = \{(5, 6), (6, 5)\}$$

$$E_{12} = \{(6, 6)\}$$

EXAMPLE 2

and we assign the value $i + j$ to the random variable X if the event E_k occurs.

$$X = \begin{cases} 2 & \text{if } E_2 \\ 3 & \text{if } E_3 \\ \vdots & \\ 11 & \text{if } E_{11} \\ 12 & \text{if } E_{12} \end{cases}$$

EXAMPLE 2

and the probability distribution associated with X is similarly defined

$$p_X(x) = P(X = x) = P(E_x) = \begin{cases} 1/36 & \text{if } X = 2 \\ 2/36 & \text{if } X = 3 \\ 3/36 & \text{if } X = 4 \\ 4/36 & \text{if } X = 5 \\ 5/36 & \text{if } X = 6 \\ 6/36 & \text{if } X = 7 \\ 5/36 & \text{if } X = 8 \\ 4/36 & \text{if } X = 9 \\ 3/36 & \text{if } X = 10 \\ 2/36 & \text{if } X = 11 \\ 1/36 & \text{if } X = 12 \end{cases}$$

The lower case x gives particular values of X , and that collection is the range of X - the values that X can take on:

$$R_X = 2, 3, 4, \dots, 10, 11, 12$$

in that language the domain of the variable X is the sample set S .

STRING LOGIC: SUMMARY



Check list for random variable

- is the sample space well defined?
- are the events mutually exclusive?
- do the events cover all the sample space?
- are the probabilities of the events defined
- are values of the variable clearly assigned one-to-one to the possible events?

(CUMULATIVE) DISTRIBUTION FUNCTION

The cumulative distribution function of a discrete random variable, X , is

$$F_X(a) = P\{X \leq a\} = \sum_{x \leq a} p_X(x)$$

it is an alternative way of describing the probabilities of different events.

Sometimes it is more convenient to use F_x than the probability function p_x .

It provided a convenient way of describing continuous random variables.

EXAMPLE

Let us imagine we have a 4 sided die.

What is the cumulative distribution function for the random variable

$X = \text{'what result appears when a 4 sided die is thrown once'}$

First we build our random variable.

All the probabilities can be assumed equally likely, and we will assign a value of the number of spots to our variable X.

We have a sample space $S=\{1,2,3,4\}$, and

$$X = \begin{cases} 1 & \text{if } \{1\} \\ 2 & \text{if } \{2\} \\ 3 & \text{if } \{3\} \\ 4 & \text{if } \{4\} \end{cases}$$

EXAMPLE

We define our probability function

$$p_X(1) = P(X = 1) = P(\{1\}) = \frac{1}{4}$$

$$p_X(2) = P(X = 2) = P(\{2\}) = \frac{1}{4}$$

$$p_X(3) = P(X = 3) = P(\{3\}) = \frac{1}{4}$$

$$p_X(4) = P(X = 4) = P(\{4\}) = \frac{1}{4}$$

or more sensibly

$$p_{X(x)} = \frac{1}{4} : x = 1, 2, 3, 4$$

We see that

$$\sum_{x \in X} p_{X(x)} = 1$$

The events that define X are exhaustive and mutually exclusive.

There is a one-to-one map between the values of the probability (mass) function. So X is a random variable, with $p_X(x)$ its probability mass function.

EXAMPLE

Using our definition of the cumulative distribution function

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

we have

$$F_X(a) = \begin{cases} 0 & a < 1 \\ 1/4 = & 1 \leq a < 2 \\ 2/4 = & 2 \leq a < 3 \\ 3/4 = & 3 \leq a < 4 \\ 1 = & a \geq 4 \end{cases}$$

Definition

$$F_X(a) = P\{X \leq a\} = \sum_{x \leq a} p_X(x) \text{ for } -\infty < a < \infty$$

if this is true, then because $p_X(x)$ obeys the axioms of probability, the following will hold:

$F_X(a)$ must be 0 as $a \rightarrow -\infty$ and 1 as $a \rightarrow \infty$

$F_X(a)$ must be monotonically increasing

$F_X(a)$ must be right continuous

CUMULATIVE DISTRIBUTION FUNCTION FOR CONTINUOUS RANDOM VARIABLES



Cumulative distributions provide a good way of describing continuous random variables.

The cumulative distribution function of a discrete random variable, X , is

$$F_X(a) = P\{X \leq a\} = \sum_{x \leq a} p_X(x)$$

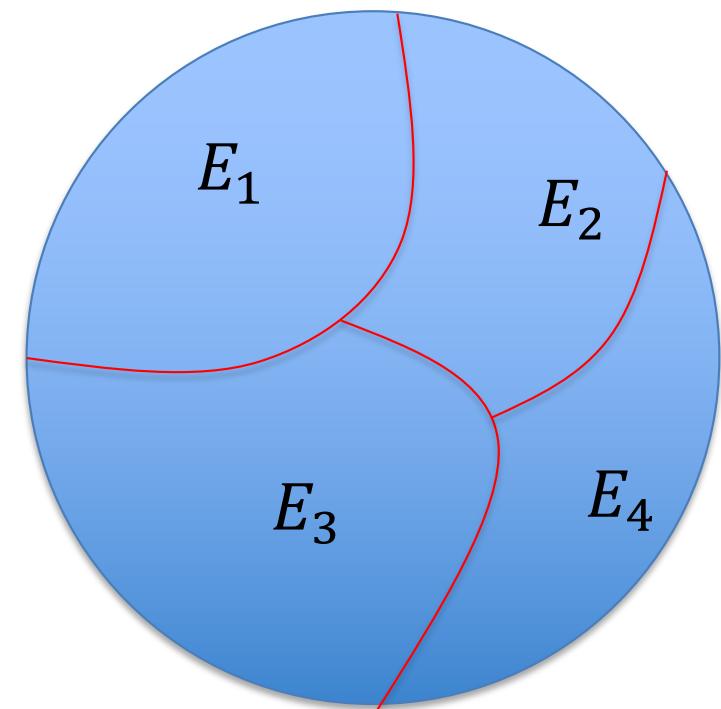
DISCRETE RANDOM VARIABLES

Definition: Consider an experiment, with outcome set S , split into n mutually exclusive and exhaustive events $E_1, E_2, E_3, \dots, E_n$. A variable, X say, which can assume exactly n numerical values each of which corresponds to one and only one of the given events is called a random variable.

Schematically the mutually exclusive and exhaustive events look like

Event types:

- Exclusive: no overlap.
- Exhaustive: all of S is covered by them.



(CUMULATIVE) DISTRIBUTION FUNCTION

The cumulative distribution function of a discrete random variable, X , is

$$F_X(a) = P\{X \leq a\} = \sum_{x \leq a} p_X(x)$$

it is an alternative way of describing the probabilities of different events.

Sometimes it is more convenient to use F_x than the probability function p_x .

It provided a convenient way of describing continuous random variables.



CONTINUOUS RANDOM VARIABLES AND CONDITIONAL PROBABILITY

INFINITE, AND CONTINUOUS PROBABILITY SPACES



Everything up to this point deals with probability spaces with a finite sample space.

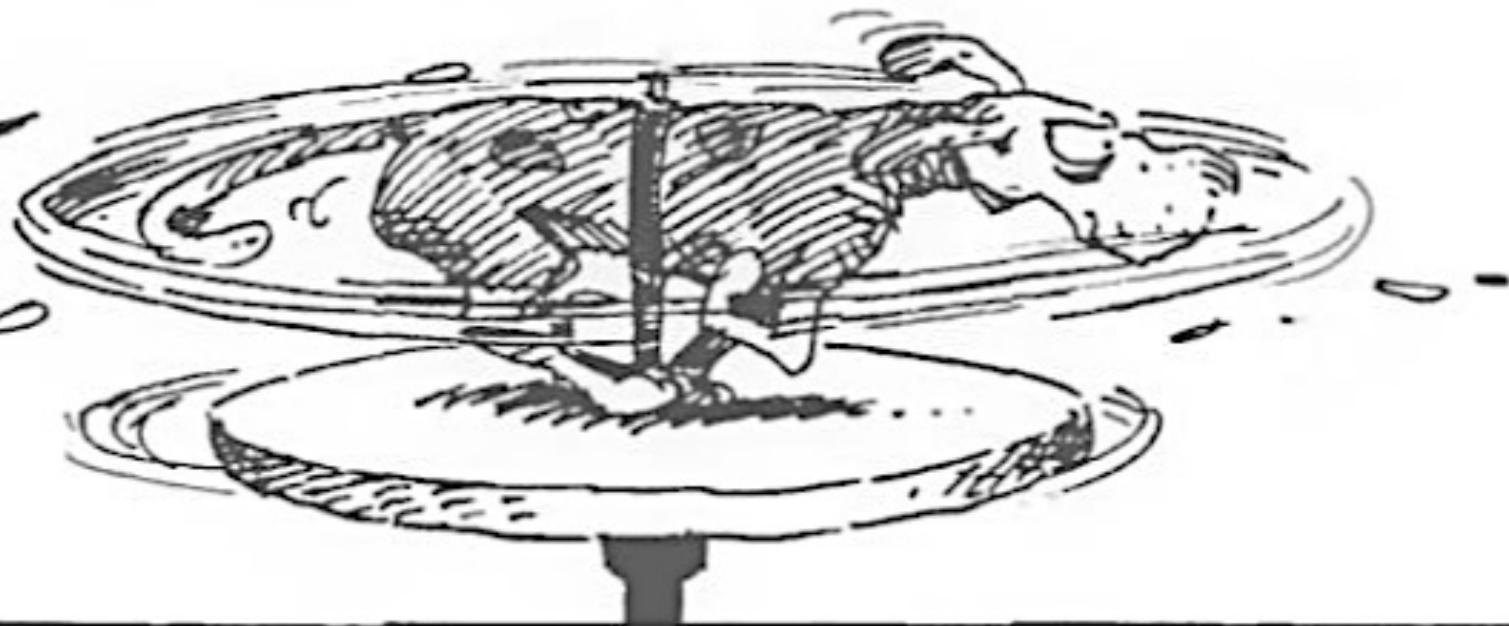
Here we mention two other cases

- when the sample space has discrete points, but an infinite number of them (all integers, for instance),
- or a prevailing situation when the sample space can take on a continuous range of values.

CONTINUOUS PROBABILITY SPACES

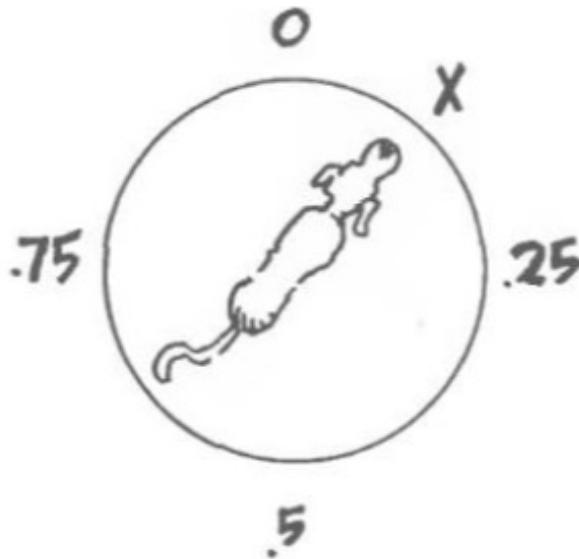
Consider a spinner - schematically a circle of unit circumference and a pointer

A SIMPLE EXAMPLE IS A BALANCED, SPINNING POINTER. IT CAN STOP ANYWHERE IN THE CIRCLE. IF X REPRESENTS THE PROPORTION OF THE TOTAL CIRCUMFERENCE IT LANDS ON, THE RANDOM VARIABLE X CAN TAKE ON ANY VALUE BETWEEN 0 AND 1—AN INFINITE RANGE OF VALUES.



CONTINUOUS PROBABILITY SPACES

If we give the spinner a whirl, the pointer will be pointing somewhere a distance x along the circumference.



It seems reasonable that every value $0 \leq x < 1$ of the distance between the pointer and the mark on the spinner is equally likely to occur. That means that the sample space is the interval $S = [0, 1)$.

We would like to have a probability model for which every value of the sample space is equally likely. We'll call the result of a spin X for now, and later we'll see that this is a *continuous random variable*.

CONTINUOUS PROBABILITY SPACES



In a similar way to before we must have

$$P(0 \leq X < 1) = 1$$

It is also the case that we expect the probability of a reading in the top half of the spinner is equal in likelihood to one in the lower half,

$$P\left(0 \leq X < \frac{1}{2}\right) = P\left(\frac{1}{2} \leq X < 1\right) = \frac{1}{2}$$

More generally, if we consider an event, $E = [a, b]$, we'd like

$$P(a \leq X < b) = b - a$$

for every a and b .

CONTINUOUS PROBABILITY SPACES

We can satisfy

$$P(a \leq X < b) = b - a$$

for every a and b for the event $E = [a, b]$ by a formula of the form

$$P(E) = \int_E f(x)dx$$

and $f(x)$ is constant function with value 1.

We call $f(x)$ *the density function* of X .

This is the generalisation of the discrete case we saw earlier:

$$P(E) = \sum_{i \in E} P(i)$$

CONDITIONAL CONTINUOUS PROBABILITIES

If we look at a process that has a density function $f(x)$, and if E is an event. We define a conditional density function by

$$f(x | E) = \begin{cases} f(x)/P(E) & \text{if } x \in E, \\ 0 & \text{if } x \notin E \end{cases}$$

Then for any event F , we have

$$P(F|E) = \int_F f(x|E)dx$$



SCHOOL OF MATHEMATICS
AND PHYSICS

MTH1005 PROBABILITY AND STATISTICS

Semester B
Lecture 5
(27/2/2024)

Danilo Roccatano

Office:

INB 3323

Email:

droccatano@lincoln.ac.uk

SUMMARY OF THE LAST LECTURE

THE KEY IDEA IS THE RANDOM VARIABLE, WHICH WE WRITE AS A LARGE



A RANDOM VARIABLE IS DEFINED AS THE NUMERICAL OUTCOME OF A RANDOM EXPERIMENT.

- Discrete random variables.
- Range and domain of the variables
- Probability (Mass/Density) Function
- Cumulative Distribution Function



CONTINUOUS RANDOM VARIABLES AND CONDITIONAL PROBABILITY

CONTINUOUS CONDITIONAL PROBABILITY



Last week, we introduced discrete probability distributions.

However similar considerations also apply to the case of continuous probabilities.

INFINITE, AND CONTINUOUS PROBABILITY SPACES



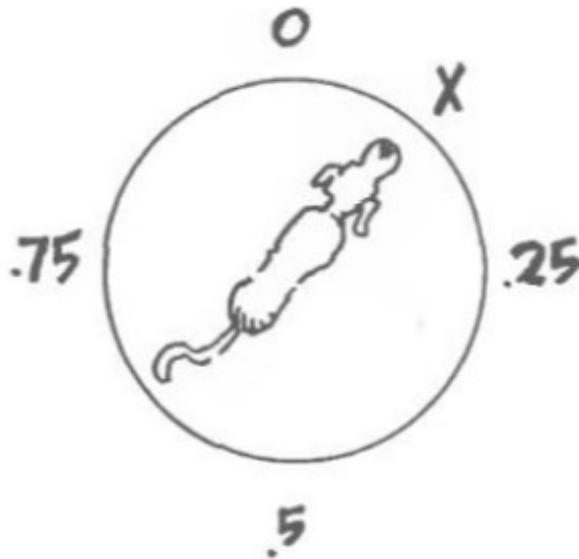
Everything up to this point deals with probability spaces with a finite sample space.

Here we mention two other cases

- when the sample space has discrete points, but an infinite number of them (all integers, for instance),
- or a prevailing situation when the sample space can take on a continuous range of values.

CONTINUOUS PROBABILITY SPACES

If we give the spinner a whirl, the pointer will be pointing somewhere a distance x along the circumference.



It seems reasonable that every value $0 \leq x < 1$ of the distance between the pointer and the mark on the spinner is equally likely to occur. That means that the sample space is the interval $S = [0, 1)$.

We would like to have a probability model for which every value of the sample space is equally likely. We'll call the result of a spin X for now, and later we'll see that this is a *continuous random variable*.

CONTINUOUS PROBABILITY SPACES



Now, we need to characterize the probability of event $E = \text{'the dog's nose pointing between points } a \text{ and } b\text{'}$ associated with the random variable X .

Since all points are equally likely, the probability is equivalent to the interval between these points along the circumference, expressed by the formula

$$P(a \leq X < b) = b - a$$

where this relationship holds true for any values of a and b within the event interval $E=[a, b]$.

CONTINUOUS PROBABILITY SPACES

The probability can alternatively be conceptualized by integrating a function, denoted as $f(x)$, over the interval of interest E , represented by the integral

$$P(E) = \int_E f(x)dx$$

where $f(x)$ is assigned constant value of 1.

This function $f(x)$ is termed the density function of X .

This is the generalisation of the discrete case we saw earlier:

$$P(E) = \sum_{i \in E} P(i)$$

CONDITIONAL CONTINUOUS PROBABILITIES

If we examine a stochastic process governed by a probability density function $f(x)$, and denote E as an event within this process, we proceed to define a conditional density function by

$$f(x | E) = \begin{cases} f(x)/P(E) & \text{if } x \in E, \\ 0 & \text{if } x \notin E \end{cases}$$

Then for any event F , we have

$$P(F|E) = \int_F f(x|E)dx$$

CONDITIONAL CONTINUOUS PROBABILITIES

We call this the conditional probability of F given E . A little manipulation makes the connection to the discrete case:

$$P(F | E) = \int_F f(x | E) dx = \int_{E \cap F} \frac{f(x)}{P(E)} dx$$

EXAMPLE OF CONDITIONAL CONTINUOUS PROBABILITY DISTRIBUTION

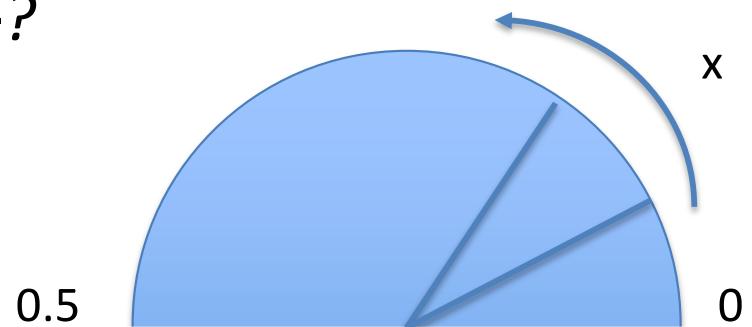
In the spinner experiment, suppose we know that the spinner has stopped with the head in the upper half of the circle, $0 \leq x \leq \frac{1}{2}$.

What is the probability that $\frac{1}{6} \leq x \leq \frac{1}{3}$?

Here

$$E = \{[0, 1/2]\},$$

$$F = \{[1/6, 1/3]\}$$



Also, we note that $F \cap E = F$

Hence

$$P(F|E) = \frac{P(F \cap E)}{P(E)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

which is reasonable, since F is $1/3$ the size of E .

EXAMPLE OF CONDITIONAL CONTINUOUS PROBABILITY DISTRIBUTION



The conditional density function here is given by

$$f(x \mid E) = \begin{cases} 2, & \text{if } 0 \leq x < 1/2, \\ 0, & \text{if } 1/2 \leq x < 1. \end{cases}$$

Thus, the conditional density function is non-zero only on $[0, 1/2]$, and is uniform there.

SUMMARY: PROBABILITY FOR CONTINUOUS PROBABILITY SPACES



We introduced the definition of continuous probability as a generalisation of the discrete case :

$$P(E) = \sum_{i \in E} P(i)$$

Using the integral instead of the summation

$$P(E) = \int_E f(x) dx$$

And the function $f(x)$ called the *probability density function* of the random variable X .

SUMMARY: CUMULATIVE DISTRIBUTION FUNCTION FOR CONTINUOUS RANDOM VARIABLES

The definition of CDF for a continuous random variable is very similar to the discrete case, but the sum is replaced by the integral, and the integrand is called the (probability) density function.

Cumulative distribution function for a continuous random variable, X

$$F_X(a) = P\{X \leq a\} = \int_{-\infty}^a f(x)dx,$$

note that conversely the fundamental theorem of calculus

$$\frac{d}{da} F(a) = f(a)$$

$F_X(a)$ must be 0 as $a \rightarrow -\infty$ and 1 as $a \rightarrow \infty$

$F_X(a)$ must be monotonically increasing

$F_X(a)$ must be right continuous

when these conditions hold, F_X defines a random variable X .

EXAMPLE

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ 1/4(y^2 + 3y) & \text{if } 0 \leq y \leq 1 \\ 1 & \text{if } y > 1 \end{cases}$$

this function fulfils all our requirements:

- as $y \rightarrow -\infty$ we see that $F_Y(y) \rightarrow 0$
- as $y \rightarrow \infty$ we see that $F_Y(y) \rightarrow 1$
- our function is right continuous - the central section is continuous (polynomial), and $F_y(0)$ and $F_y(1)$ are the same when approached from either side.

PROBABILITY DENSITY FUNCTION

We can now show that the function

$$f_Y(y) = \frac{dF_Y(y)}{dy}$$

is consistent with our previous short discussions of continuous density functions.

Being y a continuous random variable, instead of the probability mass function we have a probability density function $f(x)$ which when integrated over a valid range of X values gives the probability that the random variable X lies in that range:

$$P\{a \leq X \leq b\} = \int_a^b f_X(x)dx = F_X(b) - F_X(a)$$

the last follows from the fundamental theorem of calculus.

PROBABILITY DENSITY FUNCTION

By definition

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \lim_{\Delta y \rightarrow 0} \frac{F_Y(y + \Delta y) - F_Y(y)}{\Delta y}$$

being a little loose with what should be done with such limits we have

$$\begin{aligned} F_Y(y + \Delta y) - F_Y(y) &= f_Y(y)\Delta y \\ &= P\{y \leq Y \leq y + \Delta y\} \end{aligned}$$

so $f_Y(y)\Delta y$ is the probability that y is in the small range $[y, y + \Delta y]$.

Because we multiply $f_Y(y)\Delta y$ by a value Δy to get something sensible, it gives a naive explanation of why it is referred to as a density function.

EXAMPLE

Let the probability density function of a random variable X be

$$f(x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{2}{5}(4-x) & \text{if } 1 \leq x \leq 2 \\ 0 & \text{if } x > 2 \end{cases}$$

Suppose we want the probability that X is in the interval. From our definition we have

$$P\left\{\frac{5}{4} \leq X \leq \frac{7}{4}\right\} = \int_{\frac{5}{4}}^{\frac{7}{4}} \frac{2}{5}(4-x)dx = \frac{1}{2}.$$

We also note that the total probability, integrating over the non-zero part of $f(x)$, is unity

$$P\{-\infty \leq X \leq \infty\} = \int_1^2 \frac{2}{5}(4-x)dx = 1.$$

This means that our probability density function make sense and obey the axioms of probability.

TECHNICAL ASIDE: IMPROPER INTEGRALS

In the last example we had a integral

$$P\{-\infty \leq X \leq \infty\} = \int_{-\infty}^{\infty} f(x)dx$$

This type of integral are so-called **improper integrals** that we will encounter typically have infinite integration limits.

Similarly, to in calculus we should interpret these as a limiting process.

Typically, we will

- split up the integrals into a proper integral, and an improper integral.
- Then show that the improper integral has a well-defined limit - which we take as its value.

TECHNICAL ASIDE: IMPROPER INTEGRALS

EXAMPLE

$$f(x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{2}{5}(4-x) & \text{if } 1 \leq x \leq 2 \\ 0 & \text{if } x > 2 \end{cases}$$

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^1 0dx + \int_1^2 \frac{2}{5}(4-x)dx + \int_2^{\infty} 0dx$$

and we should take $\int_{-\infty}^1 0dx$ to mean $\lim_{a \rightarrow -\infty} \int_a^1 0dx$. As we take a to be a larger and larger negative number, the value of the integral remains 0, so this is its value.

For most/all functions we will encounter, e.g. $\int_{-\infty}^{\infty} e^{-x^2} dx$ a similar procedure yields sensible results.

Example - Uniform random variables

Let X be a random variable whose (cumulative) distribution function is

$$F_X(t) = \begin{cases} 0 & t < 0 \\ t & 0 \leq t \leq 1 \\ 1 & t > 1 \end{cases}$$

our (probability) density function is the derivative with respect to t of $F_X(t)$

$$f_X(t) = \frac{dF_X(t)}{dt} = \begin{cases} 0 & t < 0 \\ 1 & 0 \leq t \leq 1 \\ 0 & t > 1 \end{cases}$$

Since the density function is equal to 1, the area under the density over any interval between and is equal to the length of the interval.

Example - Uniform random variables

Because we have that

$$P\{a \leq X \leq b\} = \int_a^b f_X(x)dx$$

this means that the probability that the observed value for X lies in an interval contained in $(0, 1)$ is proportional to the length of the interval.

Example - Uniform random variables

To be a valid probability density function of a random variable, we must have

$$f(t) \geq 0, \text{ for all } t$$
$$\int_{-\infty}^{\infty} f(t) dt = 1$$



Properties of the random variables

Properties of a discrete random variable

Mean or Expectation value

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad \text{OR} \quad \sum_{i=1}^n \frac{x_i}{n}$$

IN THE CASE OF OUR 92 PENN STATE STUDENTS, THE MEAN WEIGHT IS

$$\sum_{i=1}^{92} \frac{x_i}{92} = \frac{13,354}{92}$$

=

145.15 POUNDS



We will see that the means is also called **expectation value of the random variable** and it is indicated ad $E[X]$

Properties of a discrete random variable

Median

FOR THE $n=92$ STUDENT WEIGHTS,
WE CAN FIND THE MEDIAN FROM THE
ORDERED STEM-AND-LEAF DIAGRAM:
JUST COUNT TO THE 46TH
OBSERVATION. THE MEDIAN IS

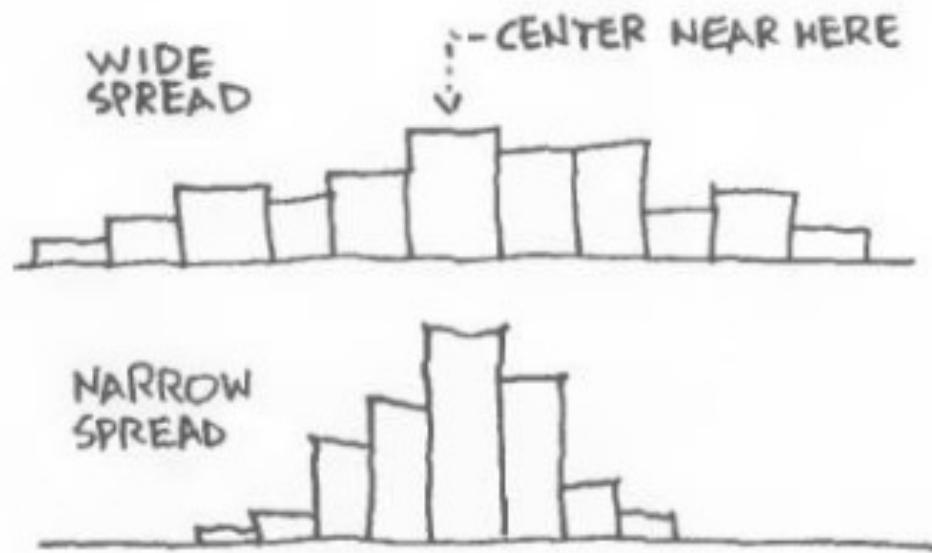
$$\frac{x_{46} + x_{47}}{2} = \frac{145 + 145}{2}$$
$$= 145 \text{ POUNDS}$$

9 : 5
10 : 288
11 : 002556688
12 : 00012355555
13 : 0000013555688
14 : 00002555558
15 : 0000000000355555555557
16 : 000045
17 : 000055
18 : 0005
19 : 00005
20 :
21 : 5

Properties of a discrete random variable

Spread

ANY SET OF MEASUREMENTS HAS TWO IMPORTANT PROPERTIES: THE CENTRAL OR TYPICAL VALUE, AND THE SPREAD ABOUT THAT VALUE. YOU CAN SEE THE IDEA IN THESE HYPOTHETICAL HISTOGRAMS.



We are going to learn how to measure these properties and how to predict that the data follow a certain trend in the distribution.

EXPECTATION OF A RANDOM VARIABLE



Now we will define the expectation and variance of a random variable. This is a further step in connecting abstract probability distributions to statistical measures and the real world, as needed.

By the end of this part, you should understand:

- How to calculate the expectation value of a random variable (this measures the central tendency of a distribution).
- How to calculate the variance of a random variable (this measures the spread or dispersion of a distribution).
- How to calculate expectation values of a function of a random variable.

Later, we'll study specific probability distributions, and understanding their expectation values will enable us to model and predict the outcomes of measurements taken on objects that can be represented as random variables.

EXPECTATION OF A RANDOM VARIABLE

Definition

Given a random variable X with a probability mass function $p(x)$ we define the expectation of X , written as $E[X]$ as

$$E[X] = \sum_{x \in X} x \cdot p(x)$$

We also call the mean of and write it as μ_X

EXAMPLE - DISCRETE RANDOM VARIABLE

Let's come back to the simulation experiment that we have done in the lab

In a game, 3 dice are rolled. The players bet £1. They gets back £1 if they rolls a single 5, £2 if 2 fives come up, and £3 if 3 fives come up (and his stake is returned).

If no 5s come up he lose their £1 stake.

EXAMPLE - DISCRETE RANDOM VARIABLE

Let us set up a random variable V for the winnings of the player.

The sample space is the cartesian product of rolling a die:

$$S = \{(x_1, x_2, x_3) : x_i = 1, 2, \dots, 6; i = 1, 2, 3\}$$

Now, we'll define V as the amount the player wins - this can be $-1, 1, 2, 3$ corresponding to 0, 1, 2, or 3 fives showing:

$$V = \begin{cases} -1 & \text{if '0 dice lands with 5 spots on the top face'} \\ 1 & \text{if '1 dice lands with 5 spots on the top face'} \\ 2 & \text{if '2 dice land with 5 spots on the top face'} \\ 3 & \text{if '3 dice land with 5 spots on the top face'} \end{cases}$$

and our probability mass function will be

$$p_V(v) = \begin{cases} \frac{125}{216} & \text{for } v = -1 \\ \frac{75}{216} & \text{for } v = 1 \\ \frac{15}{216} & \text{for } v = 2 \\ \frac{1}{216} & \text{for } v = 3 \end{cases}$$

note that $p_V(0) > 0.5$.

EXAMPLE - DISCRETE RANDOM VARIABLE

Now, suppose players play the game $n > 1$ times.

They win v_1 pounds the first time, v_2 the second time, \dots, v_n pounds the n^{th} time. The average amount one would then be a standard average

$$\bar{v} = \frac{1}{n} \sum_{i=1}^n v_i$$

Now each of the v_i can only be $-1, 1, 2, 3$.

Lets reorganise our results and say that k_{-1} times they got $v_i = -1$, k_1 times $v_i = 1$, k_2 times $v_i = 2$ and k_3 times $v_i = 3$. Where $k_{-1} + k_1 + k_2 + k_3$ will be equal to n because we are just placing our n values into these boxes, then

$$\bar{v} = \frac{1}{n} \sum_{i=1}^n v_i = (-1) \frac{k_{-1}}{n} + (1) \frac{k_1}{n} + (2) \frac{k_2}{n} + (3) \frac{k_3}{n}$$

EXAMPLE - DISCRETE RANDOM VARIABLE

$$\bar{v} = \frac{1}{n} \sum_{i=1}^n v_i = (-1) \frac{k_{-1}}{n} + (1) \frac{k_1}{n} + (2) \frac{k_2}{n} + (3) \frac{k_3}{n}$$

is our average value - but as $n \rightarrow \infty$ the ratios $\frac{k_i}{n}$ tend to our original frequentist definition of probabilities - the number of times something occurs out of the total number of attempts.

$$\frac{k_{-1}}{n} = p_V(-1), \frac{k_1}{n} = p_V(1), \frac{k_2}{n} = p_V(2), \frac{k_3}{n} = p_V(3)$$

and finally we get

$$\bar{v} = \mu_V = \sum_{i=1}^n v_i p_V(i) = \sum_{v \in R_V} v p_V(v)$$

EXAMPLE - DISCRETE RANDOM VARIABLE

in this case

$$E = \mu_V = \sum_{v \in R_V} vp_V(v) = (-1)\frac{125}{216} + (1)\frac{75}{216} + (2)\frac{15}{216} + (3)\frac{1}{216} = \frac{-17}{216} = -0.08$$

On average the player loses 8p every time they play.

SUMMARY

DEFINITION: THE **mean** OF THE RANDOM VARIABLE X IS DEFINED AS

$$\mu = \sum_{\text{all } x} xp(x)$$

MEANING:
THE CENTER
OF ITS
HISTOGRAM!



THIS IS ALSO CALLED THE EXPECTED VALUE OF X , OR $E[X]$. THINK OF IT AS THE SUM OF THE POSSIBLE VALUES, EACH WEIGHTED BY ITS PROBABILITY.

EXPECTATION OF A FUNCTION OF A DISCRETE RANDOM VARIABLE

Definitions

Let $g(X)$ be any function of a discrete random variable X . Then

$$E[g(X)] = \sum_{x \in X} g(x) \cdot p(x)$$

these can be understood in the same way as $E[X]$, it is the value of $g(X)$ times the probability of getting that value $X = x$, and hence $g(x)$.

The derivation using repetitions of experiments can be repeated, but replacing v_i with $g(v_i)$.

EXPECTATION OF A FUNCTION OF A DISCRETE RANDOM VARIABLE

In a game the player wins back a sum that is the 'square of the number of heads' on two coins.

We have a new random variable Y with probability mass function

$p(0) = \frac{1}{4}, p(1) = \frac{1}{2}, p(4) = \frac{1}{4}$ then the expectation of Y is given by

$$E[Y] = \sum_{\text{all } y} y \cdot p(y) =$$

$$0 \cdot p(0) + 1 \cdot p(1) + 4 \cdot p(4) =$$

$$0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} = \frac{3}{2} = \mu_Y$$

note that $Y = X^2$

$$E[X^2] = \sum_{\text{all } y} x^2 p(x^2)$$

$$0 \cdot p(0) + 1 \cdot p(1) + 2^2 \cdot p(2) =$$

$$0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} = \frac{3}{2} = \mu_{X^2}$$



EXPECTATION OF A FUNCTION OF A CONTINUOUS RANDOM VARIABLE

EXPECTATION OF A RANDOM VARIABLE

Definition

Given a random variable X with a probability mass function $p(x)$ we define the expectation of X , written as $E[X]$ as

$$E[X] = \sum_{x \in X} x \cdot p(x)$$

The equivalent for a continuous random variable is

$$E[Z] = \int_{-\infty}^{\infty} z \cdot f(z) dz$$

We see that the relationship is very close between discrete and continuous cases - we replace a sum with an integral.

We also call the mean of and write it as μ_X

EXAMPLE - CONTINUOUS RANDOM VARIABLE

Suppose you meet that friend of yours that is always late. You, of course, arrive on time ($T=0$) and friend shows up at a random time $T > 0$ with probability density function

$$f_T(t) = \frac{2}{15} - \frac{2t}{225}, 0 < t < 15$$

We'll ignore other times, where $f_T(t) = 0$

We could break this interval up into a large number, n , of small pieces with length

$$\Delta t = 15/n$$

$t_1, t_2, t_3, \dots, t_n$ be the midpoints of these small intervals.

EXAMPLE - CONTINUOUS RANDOM VARIABLE

The probability of your friend arriving between at time t_i is approximately

$$P(t_i - \Delta t/2 < T < t_i + \Delta t/2) \approx f_T(t_i) \Delta t.$$

We'd expect our average waiting time to be about $\sum_{i=1}^n t_i f_T(t_i) \Delta t$ - the sum of the product of length of wait, t_i , times the probability of waiting that long, $f_T(t_i) \Delta t$. The same as for the discrete random variable case.

If we now take n larger and large towards infinity, we get

$$\int_0^{15} t f_T(t) dt = \int_0^{15} t \left(\frac{2}{15} - \frac{2t}{225} \right) dt = 5 \text{ (minutes)}$$

This is the length of time you expect to wait for your friend.

SOME PROPERTIES OF E[X]

$E[a] = a$, where a is a constant for any random variable.

This is a special case of

$$E[g(X)] = \sum_{x \in X} g(x) \cdot p(x)$$

with $g(X) = a$

$$\begin{aligned} E[a] &= \sum_{x \in R_X} g(x) \cdot p_X(x) \\ &= \sum_{x \in R_X} a \cdot p_X(x) \\ &= a \sum_{x \in R_X} p_X(x) \\ &= a \end{aligned}$$

because the sum of all the $p_X(x)$ must be 1 for a valid probability mass function.

EXPECTATION OF A FUNCTION OF A RANDOM VARIABLE

Let $g(X)$ be any function of a random variable X . Then

$$E[g(X)] = \sum_{\text{all } x} g(x) \cdot p(x)$$

or for the continuous random variable Z

$$E[g(Z)] = \int_{\text{all } z} g(z) \cdot f(z) dz$$

if X is a random variable, then

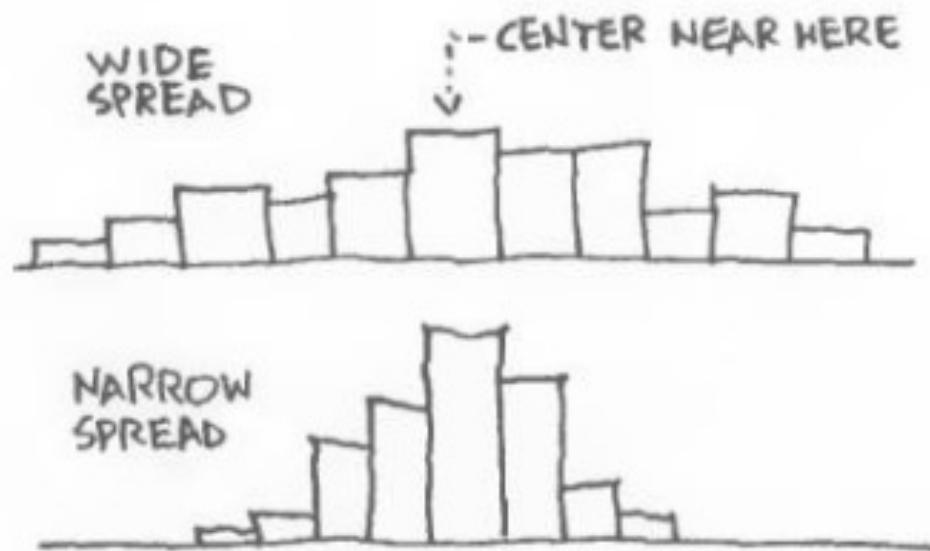
- $E[a] = a$
- $E[aX] = aE[X]$
- $E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)]$, where $g_1(X)$ and $g_2(X)$ are any functions of X .

these define the properties of a linear operator.

Properties of a discrete random variable

The next important property of random variables is the spread around the expectation value

ANY SET OF MEASUREMENTS HAS TWO IMPORTANT PROPERTIES: THE CENTRAL OR TYPICAL VALUE, AND THE SPREAD ABOUT THAT VALUE. YOU CAN SEE THE IDEA IN THESE HYPOTHETICAL HISTOGRAMS.



VARIANCE OF A RANDOM VARIABLE

The mean describes where a distribution is centred - expected value.

The **variance** describes how widely the values of the distribution are spread around the mean.

Definition

If X is a random variable with mean μ , then the variance of X , denoted by $\text{Var}(X)$ or σ_X^2 , is defined by

$$\sigma_X^2 = \text{Var}(X) = E[(X - \mu)^2]$$

this can also be written

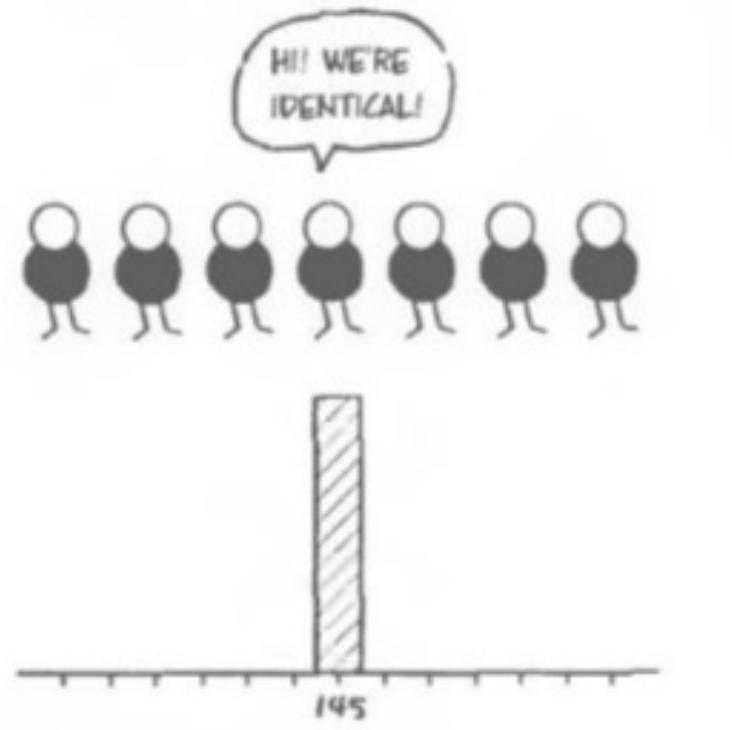
$$\sigma_X^2 = \text{Var}(X) = E[X^2] - (E[X])^2$$

It is the expected squared distance of a value from the centre of the distribution.

RANDOM VARIABLES: Measurement of the spread

MEASURES OF SPREAD

BESIDES KNOWING THE CENTRAL POINT OF A DATA SET, WE'D ALSO LIKE TO DESCRIBE THE DATA'S SPREAD, OR HOW FAR FROM THE CENTER THE DATA TEND TO RANGE. FOR INSTANCE, IF THE STUDENTS ALL WEIGHED EXACTLY 145 POUNDS, THERE WOULD BE NO SPREAD AT ALL. NUMERICALLY, THE SPREAD WOULD BE ZERO, AND THE HISTOGRAM WOULD BE SKINNY.



RANDOM VARIABLES: Measurement of the spread

BUT IF MANY OF THE STUDENTS WERE VERY LIGHT AND/OR VERY HEAVY,
OBVIOUSLY WE'D SEE SOME SPREAD—SAY, IF THE FOOTBALL TEAM WAS PART
OF THE SAMPLE...



THE HISTOGRAM WOULD BE WIDER, SOMETHING LIKE THIS:



VARIANCE OF A RANDOM VARIABLE



It is the expected value of the square of the distance of X to μ_X .

- If X only took on its average value, the variance would be 0.
- The closer the values of x are to μ the smaller the value of $\text{var}(X)$.
- The less likely values of x far from μ are, the smaller the variance.

CALCULATION OF THE VARIANCE OF A DISCRETE RANDOM VARIABLE

NOW LET'S DO THE SAME THING TO THE VARIANCE. MAYBE YOU REMEMBER THE FORMULA

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

IT (ALMOST) MEASURES THE AVERAGE SQUARED DISTANCE OF DATA FROM THE MEAN. AS ABOVE THIS CAN BE REWRITTEN:

$$s^2 = \sum_{\text{all } x} (x - \bar{x})^2 \frac{n_x}{n-1}$$



STANDARD DEVIATION

the standard deviation of a random variable X is the square root of its variance.

$$\sigma_X = \sqrt{\text{Var}(X)} = \sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2} = \sqrt{\mathbb{E}[(X - \mu)^2]}$$

a major advantage of the standard deviation is that it has the same units, if any, as the variable itself.

It gives the expected root mean squared distance of a data point from the centre of the distribution.

The smaller the variance or standard deviation, the more well defined a distribution is.

VARIANCE OF A CONTINUOUS RANDOM VARIABLE

Find $E[X^2]$ for the continuous random variable X with probability density function

$$\begin{aligned}f_X(x) &= \frac{3}{4}x(2-x) : 0 \leq x \leq 2 \\&= 0 \text{ otherwise}\end{aligned}$$

taking our general formula

$$E[h(Z)] = \int_{-\infty}^{\infty} h(z) \cdot f(z) dz$$

we fill in for our particular $f_X(x)$

$$E[X^2] = \int_0^2 x^2 \cdot \frac{3}{4}x(2-x) dx = \frac{6}{5}$$

where we ignored the parts of the improper integral $\int_{-\infty}^{\infty} x^2 \cdot f(x) dx$ where $f(x)$ is zero.

$E[g(X)]$ can be interpreted as the 'average' (mean) value of the $g(X)$. It is a number not a function.

STANDARD DEVIATION

show how the two formulae for the variance are equivalent

$$\sigma_X^2 = \text{Var}(X) = E[(X - \mu)^2] = E[X^2] - (E[X])^2$$

we can expand the bracket in $E[(X - \mu)^2]$ to get

$$E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2]$$

and we use the fact that it is a linear operator to write this as

$$E[X^2 - 2\mu X + \mu^2] = E[X^2] - E[2\mu X] + E[\mu^2]$$

and then that $E[a] = a$ and $E[aX] = aE[X]$

STANDARD DEVIATION

$$\begin{aligned} E[X^2 - 2\mu X + \mu^2] &= E[X^2] - E[2\mu X] + E[\mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - 2\mu\mu + \mu^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

because by definition $\mu = E[X]$.

USEFUL IDENTITY OF THE VARIANCE



$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

a shift of the distribution doesn't change its spread.



UNIVERSITY OF
LINCOLN

SCHOOL OF MATHEMATICS
AND PHYSICS

MTH1005

PROBABILITY AND STATISTICS

Semester B
Lecture 7 (5/3/2024)

Danilo Roccatano

Office:

INB 3323

Email:

droccatano@lincoln.ac.uk

SUMMARY OF LAST WEEK



We have given the introduced

- the expectation of random variables,
- The variance of random variables,
- Definition of moments.

We will now look at the joint random variables.

SUMMARY OF LAST WEEK

The **expectation value** gives the weighted average of the possible values of the random variable. The weighting is the likelihood that that value turns up. This is the mean of the random variable, often written as μ .

Given a random variable X with a probability mass function $p(x)$ we define the expectation of X , written as $E[X]$ as

$$E[X] = \sum_{\text{all } x} x \cdot p(x)$$

The expectation of a discrete random variable X is just the arithmetic mean of the values it takes on.

The equivalent for a continuous random variable Z is

$$E[Z] = \int_{\text{all } z} z \cdot f(z) dz$$

SUMMARY OF LAST WEEK

Definition

Let $g(X)$ be any function of a random variable X . Then

$$E[g(X)] = \sum_{x \in R_X} g(x) \cdot p(x)$$

or for the continuous random variable Z

$$E[g(Z)] = \int_{-\infty}^{\infty} g(z) \cdot f(z) dz$$

if X is a random variable, then

- $E[a] = a$
- $E[aX] = aE[X]$
- $E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)]$, where $g_1(X)$ and $g_2(X)$ are any functions of X .

these define the properties of a linear operator.

SUMMARY OF LAST WEEK: VARIANCE OF A RANDOM VARIABLE

The mean describes where a distribution is centred - expected value.

The **variance** describes how widely the values of the distribution are spread around the mean.

Definition

If X is a random variable with mean μ , then the variance of X , denoted by $\text{Var}(X)$ or σ_X^2 , is defined by

$$\sigma_X^2 = \text{Var}(X) = E[(X - \mu)^2]$$

this can also be written

$$\sigma_X^2 = \text{Var}(X) = E[X^2] - (E[X])^2$$

It is the expected squared distance of a value from the centre of the distribution.

SUMMARY OF LAST WEEK: STANDARD DEVIATION

the standard deviation of a random variable X is the square root of its variance.

$$\sigma_X = \sqrt{\text{Var}(X)} = \sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2} = \sqrt{\mathbb{E}[(X - \mu)^2]}$$

a major advantage of the standard deviation is that it has the same units, if any, as the variable itself.

It gives the expected root mean squared distance of a data point from the centre of the distribution.

The smaller the variance or standard deviation, the more well defined a distribution is.

MOMENTS OF A RANDOM VARIABLE



The mean and variance describe a probability law for a random variable to some extent - the centre of the distribution

$$\mu_X = E[X]$$

and how far points stray from the centre σ^2_X .

The k^{th} moment of a random variable X is defined as

$$m_k = E[X^k]$$

MOMENTS OF A RANDOM VARIABLE: Shape of a distribution



Therefore

$$m_1 = E[X] = \mu_X$$

And the second moment is

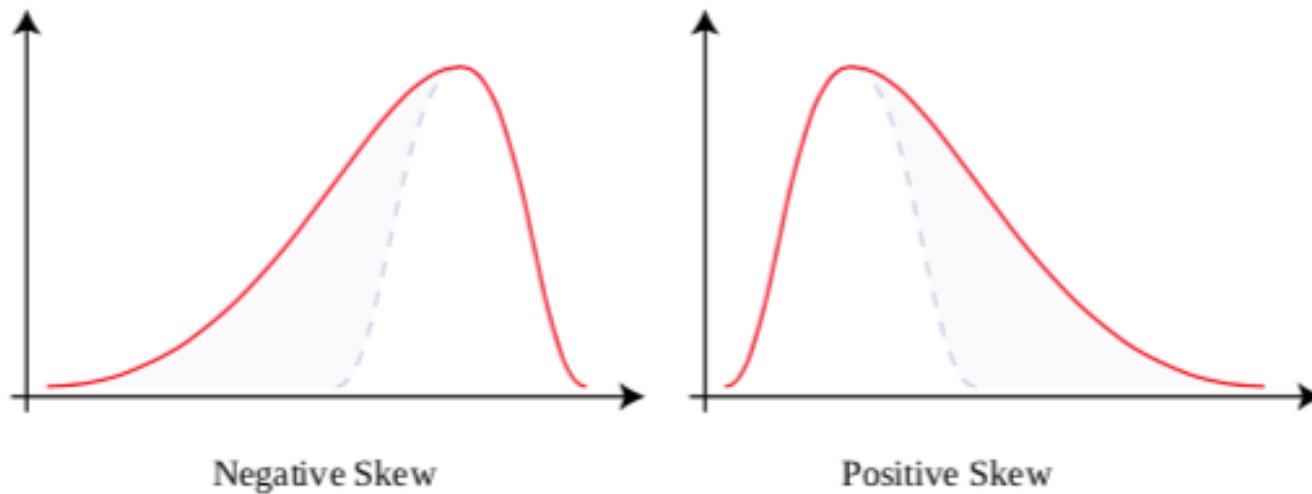
$$m_2 = E[X^2] = \sigma_X^2 + \mu_X^2$$

so the mean and variance can be defined in terms of the first two moments of the distribution.

MOMENTS OF A RANDOM VARIABLE: Shape of a distribution

Higher moments define skewness (m_3) →

- how wonky a distribution is (zero for a symmetric distribution)
- and other details of the shape of the distribution.



A full set of moments fully defines a distribution (though sometimes not all moments will be well defined).



JOINT RANDOM VARIABLES

Joint random variables refer to a concept in probability theory where we consider the behavior and outcomes of multiple random variables together.

This involves studying how these variables co-vary or interact with each other within a given probability distribution.

By analyzing joint random variables, we gain insights into the relationship between different aspects of a system or phenomenon, allowing us to make more informed predictions and decisions in various fields such as statistics, economics, and engineering.

JOINT PROBABILITY MASS FUNCTION

DEFINITION

Let X and Y be two discrete rv's defined on the sample space \mathcal{S} of an experiment. The **joint probability mass function** $p(x, y)$ is defined for each pair of numbers (x, y) by

$$p(x, y) = P(X = x \text{ and } Y = y)$$

It must be the case that $p(x, y) \geq 0$ and $\sum_x \sum_y p(x, y) = 1$.

Now let A be any particular set consisting of pairs of (x, y) values (e.g., $A = \{(x, y): x + y = 5\}$ or $\{(x, y): \max(x, y) \leq 3\}$). Then the probability $P[(X, Y) \in A]$ that the random pair (X, Y) lies in the set A is obtained by summing the joint pmf over pairs in A :

$$P[(X, Y) \in A] = \sum_{(x, y) \in A} p(x, y)$$

JOINT RANDOM VARIABLES



We define two random variables, X and Y. Consider the joint probability mass function of X and Y, denoted as

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

for discrete random variables,

It's important to note that the capitalized X represents the random variable itself, while lowercase x represents a specific value it can take on.

In the discrete case, **how many values can this function take?**
Let's examine an example.

JOINT RANDOM VARIABLES: AN EXAMPLE



Given a population of 41 students, we set up two random variables:

- $X=0$, if a student was doing pure maths, and $X = 1$ if a student was doing any other course code.
- $Y = 0$ if a student's surname did not begin with M, $Y = 1$ if a student's surname began with M.

There are 4 joint probabilities for X equal 0 or 1 and Y is 0 or 1.

So, we need to define $p(0, 0)$, $p(0, 1)$, $p(1, 0)$, $p(1, 1)$

JOINT RANDOM VARIABLES: AN EXAMPLE

Counting the students for the different cases, we can build the following Table with probabilities that both of the variables have specific values.
The joint PMF can be described using a tabular representation

	$X = 0$	$X = 1$	sum
$Y = 0$	$\frac{22}{41}$	$\frac{16}{41}$	$\frac{38}{41}$
$Y = 1$	$\frac{3}{41}$	$\frac{0}{41}$	$\frac{3}{41}$
sum	$\frac{25}{41}$	$\frac{16}{41}$	$\frac{41}{41}$

The marginal values give us the individual probability mass functions of X , $p_X(x)$ and Y , $p_Y(y)$, respectively.

JOINT RANDOM VARIABLES: MARGINAL PMF



- Once the joint pmf of the two variables X and Y is available, it is in principle straightforward to obtain the distribution of just one of these variables.

As an example, let X and Y be the number of statistics and mathematics courses, respectively, currently being taken by a randomly selected statistics major.

- Suppose that we wish the distribution of X , and that when $X = 2$, the only possible values of Y are 0, 1, and 2.

JOINT RANDOM VARIABLES: MARGINAL PMF

- Then
 - $p_X(2) = P(X = 2) = P[(X, Y) = (2, 0) \text{ or } (2, 1) \text{ or } (2, 2)]$
 $= p(2, 0) + p(2, 1) + p(2, 2)$
- That is, the joint pmf is summed over all pairs of the form $(2, y)$. More generally, for any possible value x of X , the probability $p_X(x)$ results from holding x fixed and summing the joint pmf $p(x, y)$ over all y for which the pair (x, y) has positive probability mass.

The same strategy applies to obtaining the distribution of Y by itself.

MEANINGS OF MARGINAL PMF

The marginal values along rows and columns give us the individual probability mass functions of X , $p_Y(x)$ and of Y , $p_X(y)$, respectively.

We can see this property of the marginal probability value (column or row sums) as each entry in the table is

$$P\{X = x, Y = y\} = p(x, y) = P(X = x \cap Y = y).$$

We then have using, the law of total probability

$$P(A) = \sum_{i=1}^n P(A \cap E_i) = \sum_{i=1}^n P(A | E_i)P(E_i)$$

MEANINGS OF MARGINAL PMF

we can immediately apply this result to our example for the variable X, where the E_i are the values of y , obtaining

$$P\{X = x\} = \sum_{\text{all } y} P(X = x \cap Y = y) = \sum_{\text{all } y} P(X = x \mid Y = y)P(Y = y).$$

and equally, for the Y variable, considering E_i as the values of x :

$$P\{Y = y\} = \sum_{\text{all } x} P(Y = y \cap X = x) = \sum_{\text{all } x} P(Y = y \mid X = x)P(X = x).$$

MARGINAL PMF

Definition

The marginal probability mass function of X , denoted by $p_X(x)$, is given by

$$p_X(x) = \sum_{y: p(x,y)>0} p(x,y) \quad \text{for each possible value } x$$

Similarly, the marginal probability mass function of Y is

$$p_Y(y) = \sum_{x: p(x,y)>0} p(x,y) \quad \text{for each possible value } y.$$

JOINT RANDOM VARIABLES: ANOTHER EXAMPLE



Anyone who purchases an insurance policy for a home or automobile must specify a deductible amount, the amount of loss to be absorbed by the policyholder before the insurance company begins paying out.

Suppose that a particular company offers auto deductible amounts of \$100, \$500, and \$1000, and homeowner deductible amounts of \$500, \$1000, and \$2000.

Consider randomly selecting someone who has both auto and homeowner insurance with this company and let

X = “*the amount of the auto policy deductible*”

and

Y = “*the amount of the homeowner policy deductible*”.

JOINT RANDOM VARIABLES: ANOTHER EXAMPLE

- The joint pmf of these two variables appears in the accompanying *joint probability table*:

		y		
		500	1000	5000
		.30	.05	0
x	100	.30	.05	0
	500	.15	.20	.05
	1000	.10	.10	.05

- According to this joint pmf, there are nine possible (X, Y) pairs: $(100, 500)$, $(100, 1000)$, ..., and finally $(1000, 5000)$.
- The probability of $(100, 500)$ is
$$p(100, 500) = P(X = 100, Y = 500) = .30.$$
- Clearly $p(x, y) \geq 0$, and it is easily confirmed that the sum of the nine displayed probabilities is 1.

JOINT RANDOM VARIABLES: ANOTHER EXAMPLE

- The probability $P(X = Y)$ is computed by summing $p(x, y)$ over the two (x, y) pairs for which the two deductible amounts are identical:

$$P(X = Y) = p(500, 500) + p(1000, 1000) = .15 + .10 = .25$$

		500	1000	5000
		.30	.05	0
x	100	.15	.20	.05
	500	.10	.10	.05
	1000			

- Similarly, the probability that the auto deductible amount is at least \$500 is the sum of all probabilities corresponding to (x, y) pairs for which $x \geq 500$; this is the sum of the probabilities in the bottom two rows of the joint probability table:

$$P(X \geq 500) = .15 + .20 + .05 + .10 + .10 + .05 = .65$$

		500	1000	5000
		.30	.05	0
x	100	.15	.20	.05
	500	.10	.10	.05
	1000			

JOINT RANDOM VARIABLES: ANOTHER EXAMPLE

MARGINAL PMF

The possible X values are $x = 100$, 500 , and $x = 1000$, so computing row totals in the joint probability table yields

$$p_X(100) = p(100, 500) + p(100, 1000) + p(100, 5000) = .30 + .05 + 0 = .35$$

$$p_X(500) = .15 + .20 + .05 = .40, \quad p_X(1000) = 1 - (.35 + .40) = .25$$

The marginal pmf of X is then

$$p_X(x) = \begin{cases} .35 & x = 100 \\ .40 & x = 500 \\ .25 & x = 1000 \\ 0 & \text{otherwise} \end{cases}$$

JOINT RANDOM VARIABLES: ANOTHER EXAMPLE

- Similarly, the marginal pmf of X is then

$$p_X(x) = \begin{cases} .35 & x = 100 \\ .40 & x = 500 \\ .25 & x = 1000 \\ 0 & \text{otherwise} \end{cases}$$

- From this pmf, $P(X \geq 500) = .40 + .25 = .65$, which we already calculated in Example 5.1. Similarly, the marginal pmf of Y is obtained from the column totals as

$$p_Y(y) = \begin{cases} .55 & y = 500 \\ .35 & y = 1000 \\ .10 & y = 5000 \\ 0 & \text{otherwise} \end{cases}$$



EXPECTATION OF DISCRETE JOINT RANDOM VARIABLES.

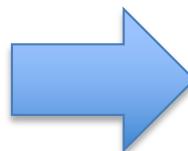
EXPECTATION OF DISCRETE JOINT RANDOM VARIABLES.

The joint **probability mass function** contains all the information about the variables it describes.

We've just seen that we can retrieve the individual probability mass/density functions by **summing** over the other variables.

If we have a general function of the variables, $g(X, Y)$, say, then we can obtain the expectation value of that function as

$$E[g(X, Y)] = \sum_y \sum_x g(x, y)p(x, y)$$



for discrete random variables and



PROBABILITY OF CONTINUOUS JOINT RANDOM VARIABLES

PROBABILITY OF CONTINUOUS JOINT RANDOM VARIABLES.



- The probability that the observed value of a continuous rv X lies in a one-dimensional set A (such as an interval) is obtained by integrating the pdf $f(x)$ over the set A .
- Similarly, the probability that the pair (X, Y) of continuous rv's falls in a two-dimensional set A (such as a rectangle) is obtained by integrating a function called the *joint density function*.

CONTINUOUS JOINT RANDOM VARIABLES.

DEFINITION

Let X and Y be continuous rv's. A joint probability density function $f(x, y)$ for these two variables is a function satisfying $f(x, y) \geq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$. Then for any two-dimensional set A

$$P[(X, Y) \in A] = \iint_A f(x, y) dx dy$$

In particular, if A is the two-dimensional rectangle $\{(x, y): a \leq x \leq b, c \leq y \leq d\}$, then

$$P[(X, Y) \in A] = P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x, y) dy dx$$

PROBABILITY OF CONTINUOUS JOINT RANDOM VARIABLES.

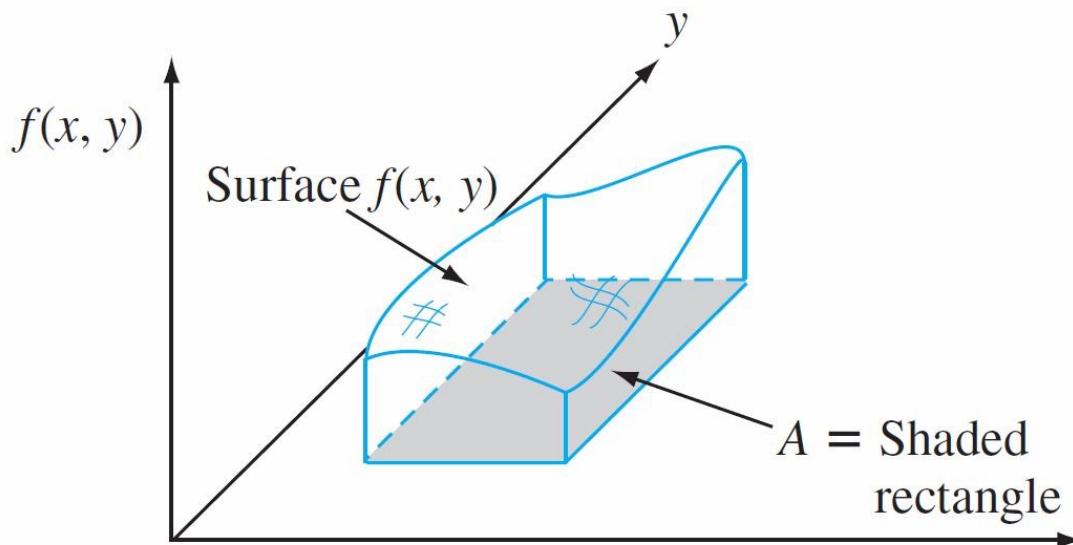


- We can think of $f(x, y)$ as specifying a surface at height $f(x, y)$ above the point (x, y) in a three-dimensional coordinate system.

Then $P[(X, Y) \in A]$ is the volume underneath this surface and above the region A , analogous to the area under a curve in the case of a single rv.

PROBABILITY OF CONTINUOUS JOINT RANDOM VARIABLES.

- This is illustrated in the following figure.



$P[(X, Y) \in A] = \text{volume under density surface above } A$

Figure 5.1

PROBABILITIES OF CONTINUOUS JOINT RANDOM VARIABLES: EXAMPLE

- A bank operates both a drive-up facility and a walk-up window. On a randomly selected day, let

X = “*the proportion of time that the drive-up facility is in use*”
(at least one customer is being served or waiting to be served)
and
 Y = “*the proportion of time that the walk-up window is in use*”.

- Then the set of possible values for (X, Y) is the rectangle

$$D = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}.$$

PROBABILITIES OF CONTINUOUS JOINT RANDOM VARIABLES: EXAMPLE

Suppose the joint pdf of (X, Y) is given by

$$f(x, y) = \begin{cases} \frac{6}{5}(x + y^2) & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = \int_0^1 \int_0^1 \frac{6}{5}(x + y^2) dx dy$$

To verify that this is a legitimate pdf, note that $f(x, y) \geq 0$ and

$$= \int_0^1 \int_0^1 \frac{6}{5} x dx dy + \int_0^1 \int_0^1 \frac{6}{5} y^2 dx dy$$

PROBABILITIES OF CONTINUOUS JOINT RANDOM VARIABLES: EXAMPLE

cont'd

$$= \int_0^1 \frac{6}{5} x \, dx + \int_0^1 \frac{6}{5} y^2 \, dy = \frac{6}{10} + \frac{6}{15} = 1$$

$$= \frac{6}{10} + \frac{6}{15}$$

$$= 1$$

The probability that neither facility is busy more than

$$P\left(0 \leq X \leq \frac{1}{4}, 0 \leq Y \leq \frac{1}{4}\right) = \int_0^{1/4} \int_0^{1/4} \frac{6}{5} (x + y^2) \, dx \, dy$$

$$= \frac{6}{5} \int_0^{1/4} \int_0^{1/4} x \, dx \, dy + \frac{6}{5} \int_0^{1/4} \int_0^{1/4} y^2 \, dx \, dy$$

PROBABILITIES OF CONTINUOUS JOINT RANDOM VARIABLES: EXAMPLE

cont'd

$$= \frac{6}{20} \cdot \left. \frac{x^2}{2} \right|_{x=0}^{x=1/4} + \frac{6}{20} \cdot \left. \frac{y^3}{3} \right|_{y=0}^{y=1/4}$$

$$= \frac{7}{640}$$

$$= .0109$$

MARGINAL PROBABILITIES OF CONTINUOUS JOINT RANDOM VARIABLE:

- The marginal pdf of each variable can be obtained in a manner analogous to what we did in the case of two discrete variables.

The marginal pdf of X at the value x results from holding x fixed in the pair (x, y) and *integrating* the joint pdf over y . Integrating the joint pdf with respect to x gives the marginal pdf of Y .

MARGINAL PROBABILITIES OF CONTINUOUS JOINT RANDOM VARIABLE

DEFINITION

The marginal probability density functions of X and Y , denoted by $f_X(x)$ and $f_Y(y)$, respectively, are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad \text{for } -\infty < x < \infty$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad \text{for } -\infty < y < \infty$$

MARGINAL PROBABILITIES OF CONTINUOUS JOINT RANDOM VARIABLES: EXAMPLE

- The marginal pdf of X , which gives the probability distribution of busy time for the drive-up facility without reference to the walk-up window, is

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_0^1 \frac{6}{5}(x + y^2) dy = \frac{6}{5}x + \frac{2}{5}$$

- for $0 \leq x \leq 1$ and 0 otherwise. The marginal pdf of Y is

$$f_Y(y) = \begin{cases} \frac{6}{5}y^2 + \frac{3}{5} & 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

MARGINAL PROBABILITIES OF CONTINUOUS JOINT RANDOM VARIABLES: EXAMPLE

- Then

$$P(.25 \leq Y \leq .75) = \int_{.25}^{.75} f_Y(y) dy = \frac{37}{80} = .4625$$

EXPECTATION OF CONTINUOUS JOINT RANDOM VARIABLES.

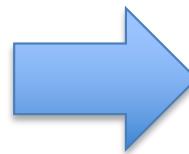
The joint **probability density function** contains all the information about the variables it describes.

$$P\{X = x, Y = y\} = f(x, y)dx dy$$

We've just seen that we can retrieve the individual probability density functions by **integrating** over the other variables.

If we have a general function of the variables, $g(X, Y)$, say, then we can obtain the expectation value of that function as

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy$$



for continuous
random variables

EXPECTATION OF SUMS OF RANDOM VARIABLES



Using the last information, we now come to a very important general result that will be the basis of a lot of the development we do later.

What we will seek to show is that

$$E[X_1 + \dots + E_n] = E[X_1] + \dots E[X_n]$$

the expectation of the sum of the random variables $X_1 \dots X_n$ is just the sum of their individual expectations.

EXPECTATION OF SUMS OF RANDOM VARIABLES

Let us start with just two variables, X and Y . We can use the general expression for the expectation of a function of two random variables

$$\begin{aligned} E[X + Y] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) dx dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y) dx dy \\ &= \int_{-\infty}^{\infty} x f_X(x) dx + \int_{-\infty}^{\infty} y f_Y(y) dy \\ &= E[X] + E[Y] \end{aligned}$$

At this point we can use proof by induction to show what we were after:

$$E[X_1 + \dots + X_n] = E[X_1] + \dots + E[X_n]$$

Example

In one of the practical, we assigned X as the 'number of heads' when two coins were flipped. We also have defined other two random variables, X_1 'the number of heads on the first coin', and X_2 'the number of heads on the second coin', so that

$$X = X_1 + X_2.$$

We have calculated the expectation value of X_1 as

$$\begin{aligned} E[X_1] &= E[X_2] = \sum_{\text{all } x_1} x_1 \cdot p(x_1) \\ &= 0 \cdot p(0) + 1 \cdot p(1) \\ &= 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} \\ &= \frac{1}{2} = \mu_{X_1} \\ &= \mu_{X_2} \end{aligned}$$

We can see that as expected

$$E[X] = E[X_1] + E[X_2] = 1$$

what about the variance of the variables?

Example

$$\begin{aligned}\mathbb{E}[X_1^2] &= \mathbb{E}[X_2^2] = \sum_{\text{all } x_1} x_1^2 \cdot p(x_1) \\&= 0 \cdot p(0) + 1 \cdot p(1) \\&= 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} \\&= \frac{1}{2}\end{aligned}$$

So

$$\text{Var}(X_1) = \text{Var}(X_2) = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

and we get in this case (because X_1 and X_2 are independent)

$$\text{Var}(X) = \text{Var}(X_1) + \text{Var}(X_2)$$

It should be clear that this would hold for tossing three coins, four coins ... n coins, which again could be proved by induction.

COVARIANCES JOINT RANDOM VARIABLES

Using the general expression of the expectation value for joint random variable (adjusted to our particular case)

$$\mathbb{E}[g(X, Y)] = \sum_y \sum_x g(x, y) p(x, y)$$

We can also write this expression as

$$\begin{aligned}\text{Cov}(X_1, X_2) &= \mathbb{E}[(X_1 - \mathbb{E}[X_1]) \cdot (X_2 - \mathbb{E}[X_2])] \\ &= \sum_{\text{all } x_1} \sum_{\text{all } x_2} (x_1 - \mu_{X_1}) \cdot (x_2 - \mu_{X_2}) \cdot p(x_1, x_2)\end{aligned}$$

USEFUL FORMULA

In a similar way to the variance the covariance can more easily be calculated using an alternative expression

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - E[X]) \cdot (Y - E[Y])] \\ &= E[XY - E[X]Y - XE[Y] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

COVARIANCES JOINT RANDOM VARIABLES

So,

$$\begin{aligned}\text{Cov}(X_1, X_2) &= E\left[(X_1 - E[X_1]) \cdot (X_2 - E[X_2])\right] \\ &= \sum_{\text{all } x_1} \sum_{\text{all } x_2} (x_1 - \mu_{X_1}) \cdot (x_2 - \mu_{X_2}) \cdot p(x_1, x_2)\end{aligned}$$

In this case we can very reasonably expect that

$$p(x_1, x_2) = p(x_1)p(x_2)$$

(remember the definition of independent probabilities).

EXAMPLE

Calculate $E[X_1 X_2]$ and $\text{Cov}(X_1, X_2)$ for our two coins example.

$$E[X_1 X_2] = \sum_{\text{all } x_1} \sum_{\text{all } x_2} (x_1) \cdot (x_2) \cdot p(x_1, x_2) =$$

$$\sum_{\text{all } x_1} \sum_{\text{all } x_2} (x_1) \cdot (x_2) \cdot p(x_1) \cdot p(x_2) =$$

$$(0) \times (0) \times \frac{1}{2} \times \frac{1}{2} +$$

$$(1) \times (0) \times \frac{1}{2} \times \frac{1}{2} +$$

$$(0) \times (1) \times \frac{1}{2} \times \frac{1}{2} +$$

$$(1) \times (1) \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

COVARIANCES JOINT RANDOM VARIABLES

We can now write out all the terms in the double sum:

$$\begin{aligned} \sum_{\text{all } x_1} \sum_{\text{all } x_2} (x_1 - \mu_{X_1}) \cdot (x_2 - \mu_{X_2}) \cdot p(x_1, x_2) = \\ (0 - \frac{1}{2}) \times (0 - \frac{1}{2}) \times \frac{1}{2} \times \frac{1}{2} + \\ (1 - \frac{1}{2}) \times (0 - \frac{1}{2}) \times \frac{1}{2} \times \frac{1}{2} + \\ (0 - \frac{1}{2}) \times (1 - \frac{1}{2}) \times \frac{1}{2} \times \frac{1}{2} + \\ (1 - \frac{1}{2}) \times (1 - \frac{1}{2}) \times \frac{1}{2} \times \frac{1}{2} = \\ \frac{1}{16} - \frac{1}{16} - \frac{1}{16} + \frac{1}{16} = 0 \end{aligned}$$

this will be the case whenever the two variables are independent.

EXAMPLE

looking back we see that this is indeed equal to $E[X_1]E[X_2]$, so again

$$\begin{aligned}\text{Cov}(X_1, X_1) &= E\left[(X_1 - E[X_1]) \cdot (X_2 - E[X_2])\right] \\ &= E[X_1 X_2] - E[X_1]E[X_2] = 0\end{aligned}$$

SUMMARY: COVARIANCES JOINT RANDOM VARIABLES

- Random variables whose covariance is zero are called uncorrelated $\text{Cov}(X,Y)=0$
- In the case that the two variables are independent the covariance is also zero as

$$p(x_1, x_2) = p(x_1)p(x_2)$$

and

$$\begin{aligned} E[X_1 X_2] &= E[X_1] E[X_2] \\ \Rightarrow \text{Cov}(X, Y) &= E[XY] - E[X]E[Y] = 0 \end{aligned}$$

However, note that the $\text{Cov}(X,Y)$ can be zero even if the two variable are dependent.

MEANING OF COVARIANCE

It indicates whether there is a relationship between two random variables. We just saw that in the case of two independent random variables the covariance will be zero.

Let's define two special random variables, X and Y. They indicate whether or not events A and B occur. They are sometimes called **indicator variables**.
In our language of random variables

if A occurs

$$X = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$
$$Y = \begin{cases} 1 & \text{if } B \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

and we can then see that

$$XY = \begin{cases} 1 & \text{if } X = 1, Y = 1 \\ 0 & \text{otherwise} \end{cases}$$

MEANING OF COVARIANCE



So if we look at the covariance of X and Y we get

$$\begin{aligned} \text{Cov}(X, Y) &= E[XY] - E[X]E[Y] \\ &= P(\{X = 1, Y = 1\}) - P(\{X = 1\})P(\{Y = 1\}) \end{aligned}$$

the last line comes from the definition of the variables, it is the proportion that the variable is in the 1 state.

All the other terms in the general expression are zero because we multiply by zero.

MEANING OF COVARIANCE

What if the covariance was greater than 0?

What would that imply for the variables (and therefore the events A and B)?

$$\text{Cov}(X, Y) > 0 \implies P(\{X = 1, Y = 1\}) > P(\{X = 1\})P(\{Y = 1\})$$

$$\implies \frac{P(\{X = 1, Y = 1\})}{P(\{X = 1\})} > P(\{Y = 1\})$$

Using the definition of a conditional probability

$$\implies P(Y = 1 | X = 1) > P(Y = 1)$$

This means that if the event A has happened (so $X=1$), it becomes more likely that event B ($Y=1$) will also happen.

Note that we could swap X and Y above, and it would imply that B became more likely when A has occurred too.

MEANING OF COVARIANCE



It might be that the covariance is negative. Then we get

$$\text{Cov}(X, Y) < 0 \implies P(\{X = 1, Y = 1\}) < P(\{X = 1\})P(\{Y = 1\})$$

$$\implies \frac{P(\{X = 1, Y = 1\})}{P(\{X = 1\})} < P(\{Y = 1\})$$

$$\implies P(Y = 1 | X = 1) < P(Y = 1)$$

So, in this case, the event A happening makes B less likely.

CORRELATION

In general it can be shown that a positive $\text{Cov}(X,Y)$ is an indication that Y increases when X does. A negative $\text{Cov}(X,Y)$ is an indication that Y decreases when X increases.

What we want is a better indicator than the covariance.

The difficulty with the covariance is that it has dimensions, and its size is dependent on the definition of the variables. Instead, or in addition, we can use the correlation. definition

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

CORRELATION

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

It can be shown that the correlation will always have a value between -1 and $+1$.

The significance of the correlation is similar to just discussed for the covariance:

- Positive correlation between two variables means that X increases as Y increases.
- Negative correlation means that X decreases as Y increases.



SCHOOL OF MATHEMATICS
AND PHYSICS

MTH1005 PROBABILITY AND STATISTICS

*Semester B
Lecture 8 (12/3/2024)*

Danilo Roccatano

Office:

INB 3323

Email:

droccatano@lincoln.ac.uk

SUMMARY LAST LESSON



We have

- defined the covariance of two random variables X and Y
- examined the meaning of the covariance
- shown that $\text{Cov}(X, Y) = 0$ for independent random variables.
- defined the correlation between two random variables X and Y .



The Simple Linear Regression Model

SIMPLE LINEAR REGRESSION MODEL



- The simplest deterministic mathematical relationship between two variables x and y is a linear relationship

$$y = a_0 + a_1 x$$

- The set of pairs (x, y) for which $y = a_0 + a_1 x$ determines a straight line with slope a_1 and y -intercept a_0 . The objective of this section is to develop a linear probabilistic model.
- If the two variables are not deterministically related, then for a fixed value of x , there is uncertainty in the value of the second variable.

SIMPLE LINEAR REGRESSION MODEL



- More generally, the variable whose value is fixed by the experimenter will be denoted by x and will be called the **independent, predictor, or explanatory variable**.
- For fixed x , the second variable will be random; we denote this random variable and its observed value by Y and y , respectively, and refer to it as the **dependent or response variable**.
- Usually observations will be made for a number of settings of the independent variable.

SIMPLE LINEAR REGRESSION MODEL

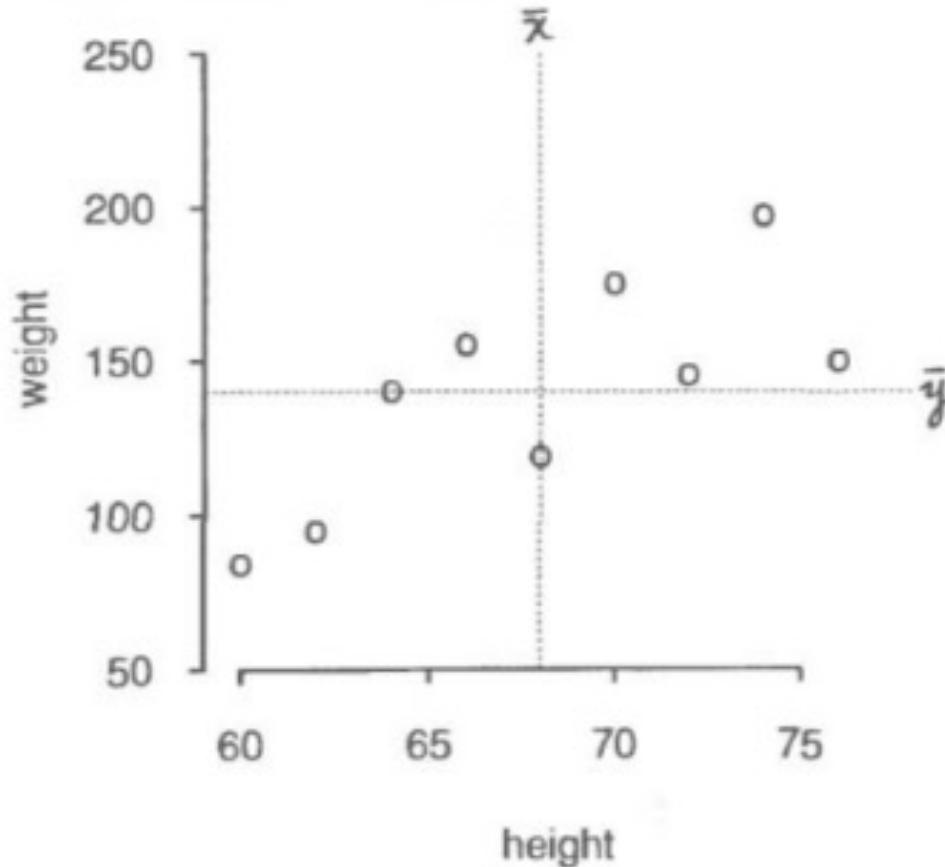


- Let x_1, x_2, \dots, x_n denote values of the independent variable for which observations are made, and let Y_i and y_i , respectively, denote the random variable and observed value associated with x_i . The available bivariate data then consists of the n pairs $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.
- A picture of this data called a **scatter plot** gives preliminary impressions about the nature of any relationship. In such a plot, each (x_i, y_i) is represented as a point plotted on a two-dimensional coordinate system.

SIMPLE LINEAR REGRESSION MODEL

We want to correlate the weight with the height of a sample of students

HEIGHT	WEIGHT
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150



SIMPLE LINEAR REGRESSION MODEL



Given our assumption of a straight-line

$$y_i = a_0 + a_1 x_i + e_i$$

the error at each point is given by

$$e_i = y_i - a_0 - a_1 x_i$$

So, in some sense the sum of the total errors would be given by the sum of the errors. We will take the sum of the squares of the errors

$$S_r = \sum_{i=0}^{n-1} e_i^2 = \sum_{i=0}^{n-1} (y_i - a_0 - a_1 x_i)^2$$

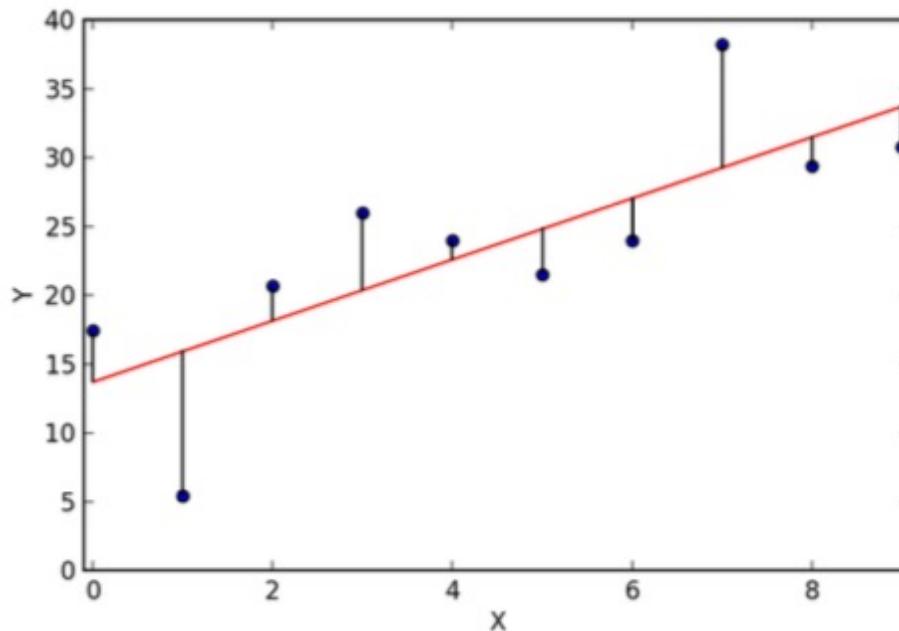
as our error criterion.

SIMPLE LINEAR REGRESSION MODEL

So in some sense the sum of the total errors would be given by the sum of the errors. We will take the sum of the squares of the errors

$$S_r = \sum_{i=0}^{n-1} e_i^2 = \sum_{i=0}^{n-1} (y_i - a_0 - a_1 x_i)^2$$

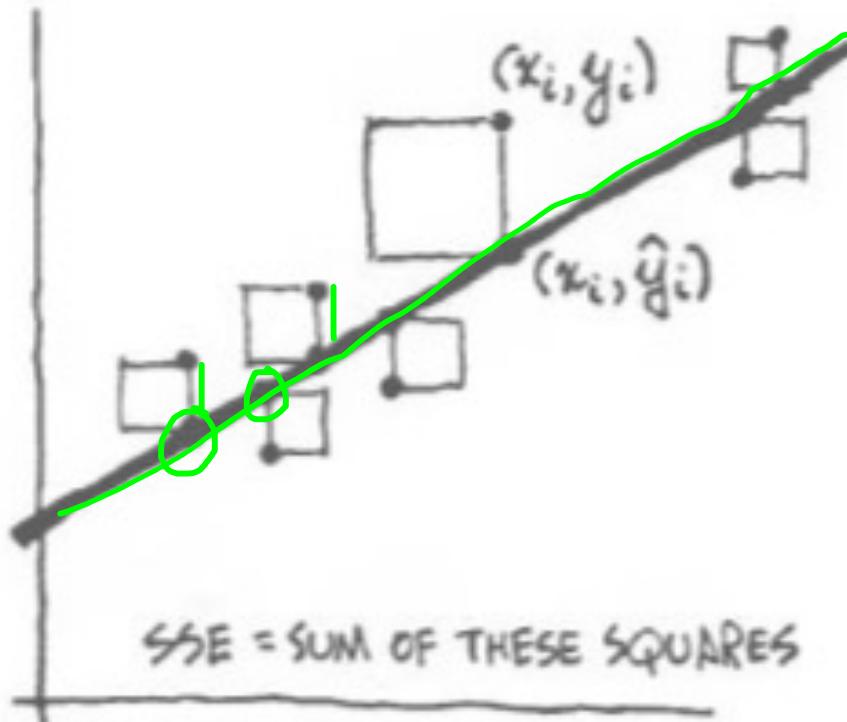
as our error criterion.



SIMPLE LINEAR REGRESSION MODEL

THE IDEA IS TO MINIMIZE THE TOTAL SPREAD OF THE y VALUES FROM THE LINE. JUST AS WHEN WE DEFINED THE VARIANCE, WE LOOK AT ALL THE SQUARED y DISTANCES FROM THE LINE, AND ADD THEM UP TO GET THE SUM OF SQUARED ERRORS:

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



IT'S AN AGGREGATE MEASURE OF HOW MUCH THE LINE'S "PREDICTED y_i ," OR \hat{y}_i , DIFFER FROM THE ACTUAL DATA VALUES y_i .

DETERMINATION OF THE OPTIMAL PARAMETERS

If we look at our model

$$S_r = \sum_{i=0}^{n-1} e_i^2 = \sum_{i=0}^{n-1} (y_i - a_0 - a_1 x_i)^2$$

we see that there are 2 parameters, a_0 and a_1 that control the slope and intercept of our model.

It is a model, we are assuming that there is a linear relationship between x and y.

We want to minimize the value of S_r , so we differentiate with respect to our parameters

DETERMINATION OF THE OPTIMAL PARAMETERS

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_i)$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum [(y_i - a_0 - a_1 x_i) x_i]$$

where the summations go from 0 to n - 1 .

Note that the points (x_i, y_i) are not variables, they are things we have measured. What we can vary is the parameters of our model. So S_r is a function of the two parameters a_0 and a_1 .

For more general models we will have more parameters and a more complex relationship than the straight line assumed here.

DETERMINATION OF THE OPTIMAL PARAMETERS

We want to minimize the value of S_r , so we differentiate with respect to our parameters

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_i)$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum [(y_i - a_0 - a_1 x_i) x_i]$$

This gives us a pair of simultaneous linear equations, sometimes called the normal equations.

We can solve these for a_1 and a_0 .

$$a_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

and plug this into the first equation to get

$$a_0 = \frac{\sum y_i}{n} - a_1 \frac{\sum x_i}{n} = \bar{y} - a_1 \bar{x}$$

where \bar{y} and \bar{x} are the means of the x and y values.

DET. OF THE OPTIMAL PARAMETERS: THE EXAMPLE

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
60	84	-8	-56	64	3136	448
62	95	-6	-45	36	2025	270
64	140	-4	0	16	0	0
66	155	-2	15	4	225	-30
68	119	0	-21	0	441	0
70	175	2	35	4	1225	70
72	145	4	5	16	25	20
74	197	6	57	36	3249	342
76	150	8	10	64	100	80
SUM=612	1260			$SS_{xx} = 240$	$SS_{yy} = 10426$	$SS_{xy} = 1200$
$\bar{x} = 68$ $\bar{y} = 140$						

OPTIMAL PARAMETERS: THE EXAMPLE

WHICH GIVES VALUES OF a AND b :

$$b = \frac{1200}{240} = 5 \quad a = \bar{y} - b\bar{x} = 140 - 5(68) = -200$$

so $y = -200 + 5x$

$$y = a + bx$$

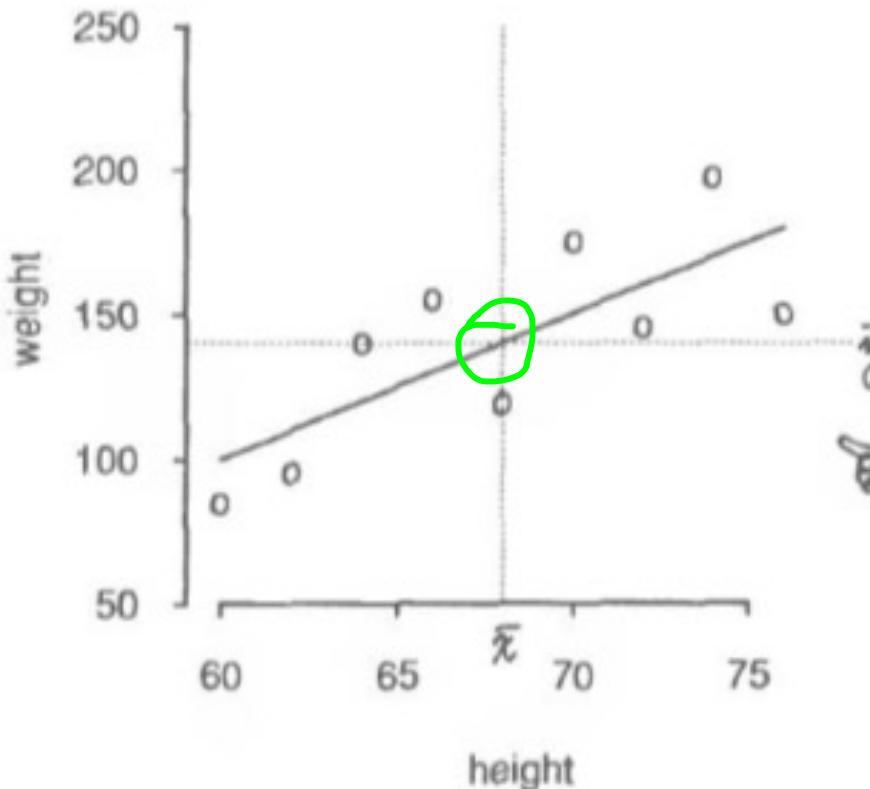
WHERE

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

AND

$$a = \bar{y} - b\bar{x}$$

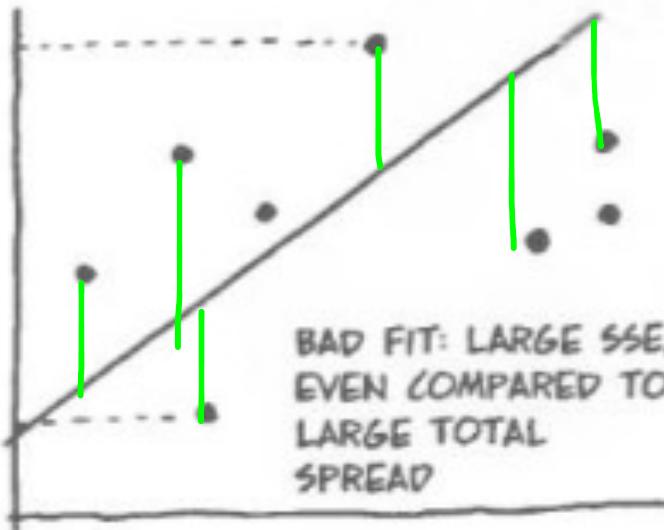
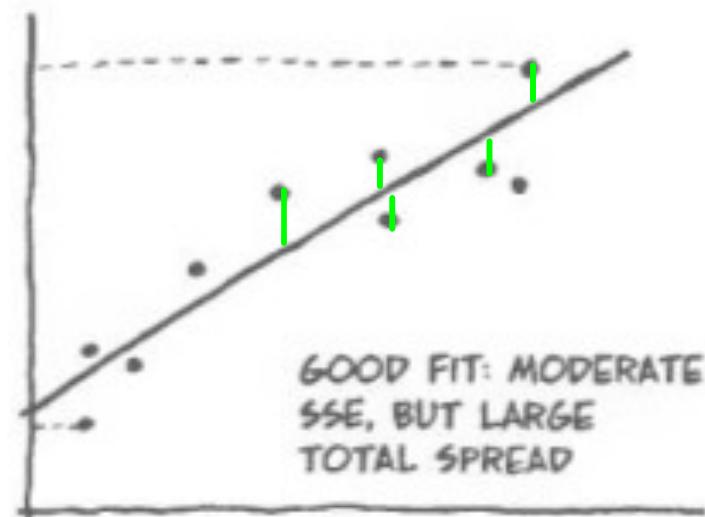
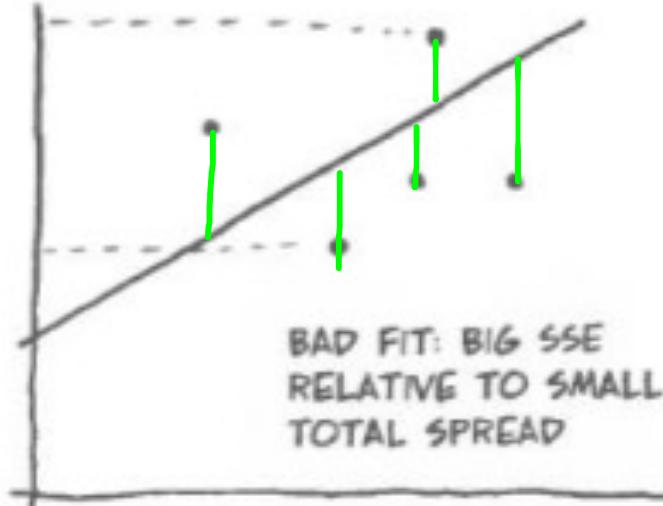
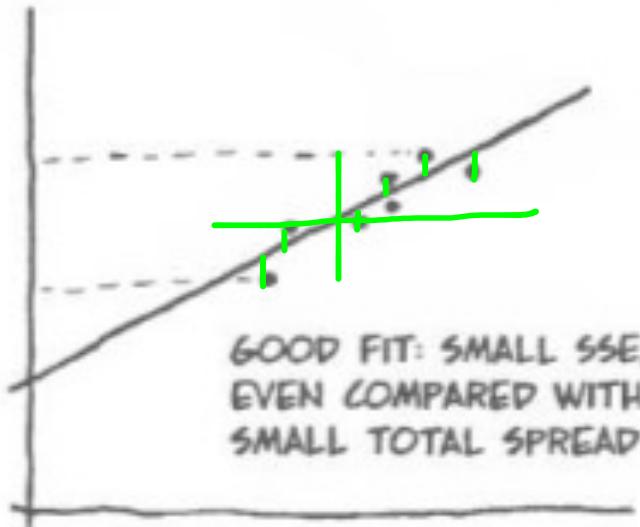
(HERE \bar{x} AND \bar{y} ARE THE MEANS OF $\{x_i\}$ AND $\{y_i\}$ RESPECTIVELY.)



NOTE:
THE REGRESSION
LINE ALWAYS
PASSES THROUGH
THE POINT
 (\bar{x}, \bar{y}) !!!



QUANTIFYING THE ERROR



QUANTIFYING THE ERROR

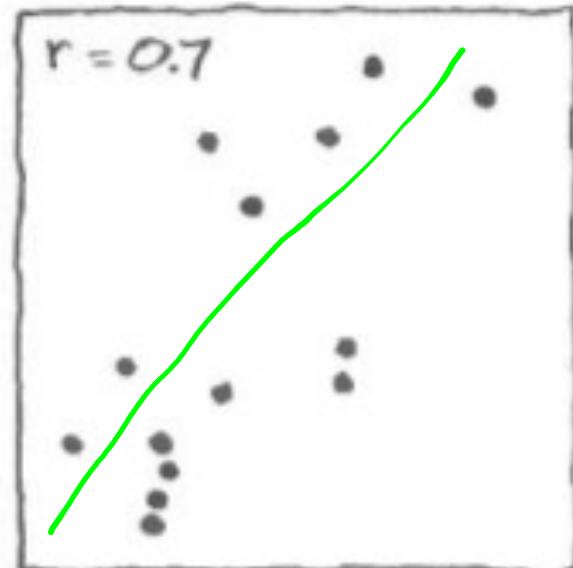
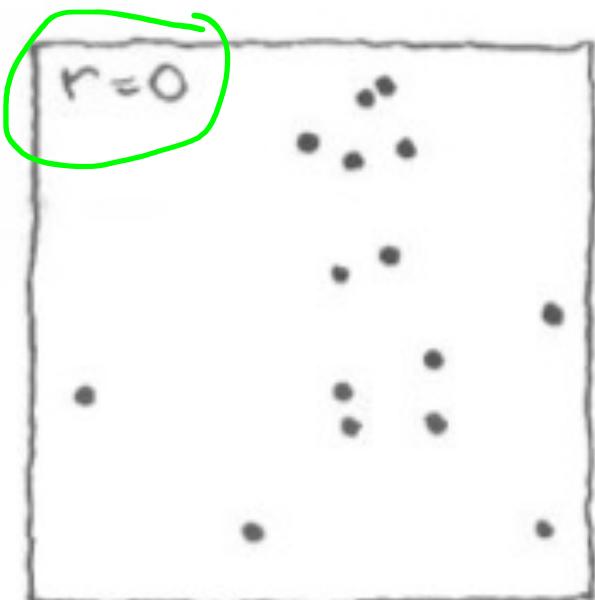
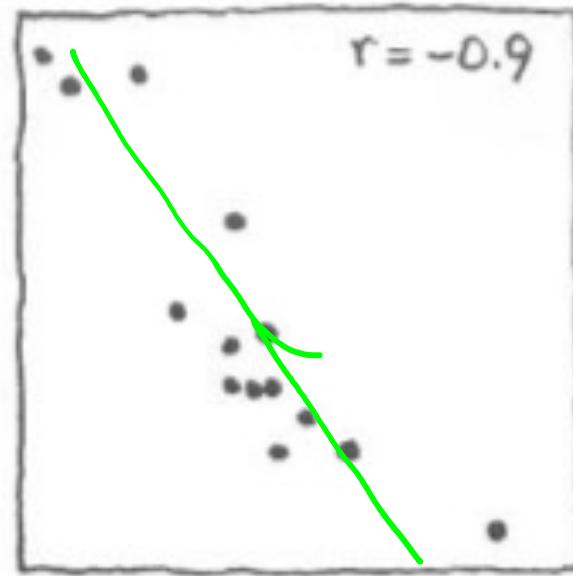
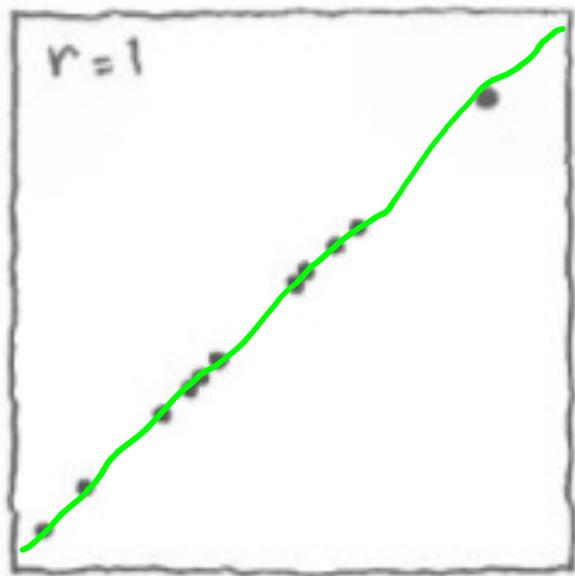
To quantify the error in our model we can look at the **correlation coefficient (r)** of our data, conventionally called **r**.

This is the covariance of the data (x, y) divided by the square roots of the variance of x and y.

$$r = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

The measures the difference between the scatter of the data from the mean and the scatter of the data from the ideal straight line.

QUANTIFYING THE ERROR



MEASURES OF CENTRAL TENDENCY - REVIEW

Where is the middle of the distribution?

We have looked at the mean, which we saw is also termed the expectation value of a random variable.

There are two other common measures of the centre of the distribution

- **Mean** *of data or expectation value of a random variable*
- **Mode**, *which is the most commonly occurring value in the distribution (may not be unique)*
- **Median**, *which is middle value of the distribution if values are listed in order, for an odd number of values. Or the mean of the two central values, for even numbers of data points.*

These measures are often applied to samples, which may or may not come from some idealised probability distribution.

RANDOM VARIABLES: Properties of the random variable

Mean or Expectation value

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad \text{OR} \quad \sum_{i=1}^n \frac{x_i}{n}$$

IN THE CASE OF OUR 92 PENN STATE STUDENTS, THE MEAN WEIGHT IS

$$\sum_{i=1}^{92} \frac{x_i}{92} = \frac{13,354}{92}$$

=

145.15 POUNDS



We will see that the mean is also called **expectation value of the random variable** and it is indicated as $E[X]$

MEDIAN FOR DISCRETE DATA

For a discrete set of data the median is easily defined:

median is middle value of the distribution: if values are listed in order, for an odd number of values. Or the (arithmetic) mean of the two central values, for even numbers of data points.

$$(1, 2, 3, 4, 5)$$

has median 3, the central value.

$$(1, 2, \textcircled{3}, 4, 5, 6)$$

has median 3.5 - the mean of 3 and 4.

Algorithmically, we sort the list of numbers, check whether there are an odd or even number, then select the central value, or the central two and average. We could also think of this as defining a partition that cuts the data in to two pieces, of equal weight i.e. with equal numbers of data in each partition.

RANDOM VARIABLES: Properties of the random variable

Median

FOR THE $n=92$ STUDENT WEIGHTS,
WE CAN FIND THE MEDIAN FROM THE
ORDERED STEM-AND-LEAF DIAGRAM:
JUST COUNT TO THE 46TH
OBSERVATION. THE MEDIAN IS

$$\frac{x_{46} + x_{47}}{2} = \frac{145 + 145}{2}$$
$$= 145 \text{ POUNDS}$$

9 : 5
10 : 288
11 : 002556688
12 : 00012355555
13 : 0000013555688
14 : 00002555558
15 : 0000000000355555555557
16 : 000045
17 : 000055
18 : 0005
19 : 00005
20 :
21 : 5

MEDIAN OF A DISCRETE RANDOM VARIABLE

We could also think of this as defining a partition that cuts the data in to two pieces, of equal weight i.e. with equal numbers of data in each partition.

The median is the value where it is likely that the variable would take a value greater than or equal, i.e. $P(Z \geq z_{\text{median}}) \geq 0.5$, or less than or equal to , i.e. $P(Z \leq z_{\text{median}}) \geq 0.5$.

Remember that the cumulative distribution function of a discrete random variable Z is given by

$$F_Z(a) = P(Z \leq a) = \sum_{z \leq a} p_Z(z)$$

If we look back at the list (1, 2, 3, 4, 5) and convert it to a random variable - say that each value is equally likely, we get

$$p_Z = \frac{1}{5}, \quad z = 1,2,3,4,5$$

MEDIAN OF A DISCRETE RANDOM VARIABLE

Now

$$F_Z(z) = \begin{cases} 0, & z < 1 \\ \frac{1}{5}, & 1 \leq z < 2 \\ \frac{2}{5}, & 2 \leq z < 3 \\ \frac{3}{5}, & 3 \leq z < 4 \\ \frac{4}{5}, & 4 \leq z < 5 \\ 1, & z \geq 5 \end{cases}$$

We can deduce that the median must be 3:

$$P(Z \leq 3) = F(3) = \frac{3}{5}.$$

The other equality is a bit more tricky

$$P(Z \geq 3) = \sum_{z \geq 3} p_Z(z) = p_Z(3) + p_Z(4) + p_Z(5) = \frac{3}{5}$$

MEDIAN FOR CONTINUOUS DATA

The computation of the **median for a continuous random variable** is more interesting.

Recall cumulative distribution function of a random variable gives the probability that the random variable will take on a value equal to, or smaller than, the argument of the function. For a continuous random variable X

$$F_X(a) = \int_{-\infty}^a f(x)dx = P(X \leq a)$$

If we take a value a such that $F_X(a) = 0.5$ this means that at that value the probability

$$P(X \leq a) = P(X > a) = \frac{1}{2}$$

If we think about it this would also apply to the discrete case. If we take a random value from our list, half the time it will be smaller than our median, and half the time it will be larger.

From this argument, we can say that a is the median of the random variable X .

SUMMARY OF CALCULATING THE MEDIAN



Median of a discrete set of data is the middle value of the distribution if values are listed in order, for an odd number of values. Or the mean of the two central values, for even numbers of data points.

The median of a discrete random variable Z is the value z such that

$$P(Z \leq z) \geq 0.5$$

The median of a continuous random variable X is the value x such that

$$F(x) = \frac{1}{2}.$$

MEDIAN, QUANTILES AND STATISTICAL TABLES

We start to examine in more detail the properties of a probability distribution, and hence the distribution of the values of a random variable. Then we looked at

- Measures of Central tendency
 - Mean, Mode
 - Medians

We now look to other statistics descriptors to measure the variation

- Quantiles and Percentiles

as well as a remarkable result about probability distributions called
=> **Chebyshev's Inequality.**

which leads to

=> **The Weak Law of Large Numbers**

MEASURES OF VARIATION

We have already looked at how to calculate the variance and standard deviation of a random variable.

They measure how much the different values of a random variable differ from the mean of the variable, and how likely the given value is to occur.

- **Variance and standard deviation**
- **Range:** *largest value minus the smallest value.*
- **inter-quartile distance:** *the 'distance' between the first quartile and the third quartile*

These measures are often applied to samples, which may or may not come from some idealised probability distribution.

The first two we have covered, or are obvious. What are quartiles?

QUANTILES

Quantiles are cut points dividing a set of observations into equal sized groups.

Actually, we've already met one - the median. This cuts the set of observations/values of the distribution in 2.

For a continuous distribution life is actually easier, the median is just the value at which the cumulative distribution function of a variable reaches $\frac{1}{2}$.

- median of a discrete set of data is the middle value of the distribution if values are listed in order, for an odd number of values. Or the mean of the two central values, for even numbers of data points.
- the median of a discrete random variable Z is the value z such that $P(Z \leq z) \geq 0.5$ and $P(Z \geq z) \geq 0.5$.
- median of a continuous random variable X is the value x such that $F(x) = \frac{1}{2}$.

QUANTILES



For a general quantile we just replace

$$\frac{1}{2} \text{ with } \frac{a}{n}$$

where $(n - 1)$ is number of cuts we will make in the data, gives a^{th} n-tile.

Some of them have special names.

QUARTILES



If we take $n=4$, then we would make 3 cuts in the data - we've quartered it.

the values at which we make the three cuts are called quartiles:

- **first quartile**, Q_1 , is defined by $F(Q_1) = \frac{1}{4}$
- **second quartile**, Q_2 , is defined by $F(Q_2) = \frac{2}{4} = \frac{1}{2}$
- **third quartile**, Q_3 , is defined by $F(Q_3) = \frac{3}{4}$

Note again that Q_2 is the median.

EXAMPLE

Consider a probability density function for the random variable Y

$$f_Y(y) = \begin{cases} 2e^{-2y} & y > 0, \\ 0 & \text{otherwise} \end{cases}$$

The cumulative distribution function is

$$F_Y(a) = \int_0^a f(y)dy = 1 - e^{-2a}$$

we can then calculate any quartile as $F_Y(Q_n) = \frac{n}{4}$

This implies, making $F_Y(Q_n)$ explicit, that

$$1 - e^{-2Q_n} = \frac{n}{4}$$

and rearranging we get

$$Q_n = -\frac{\ln(1 - \frac{n}{4})}{2}$$

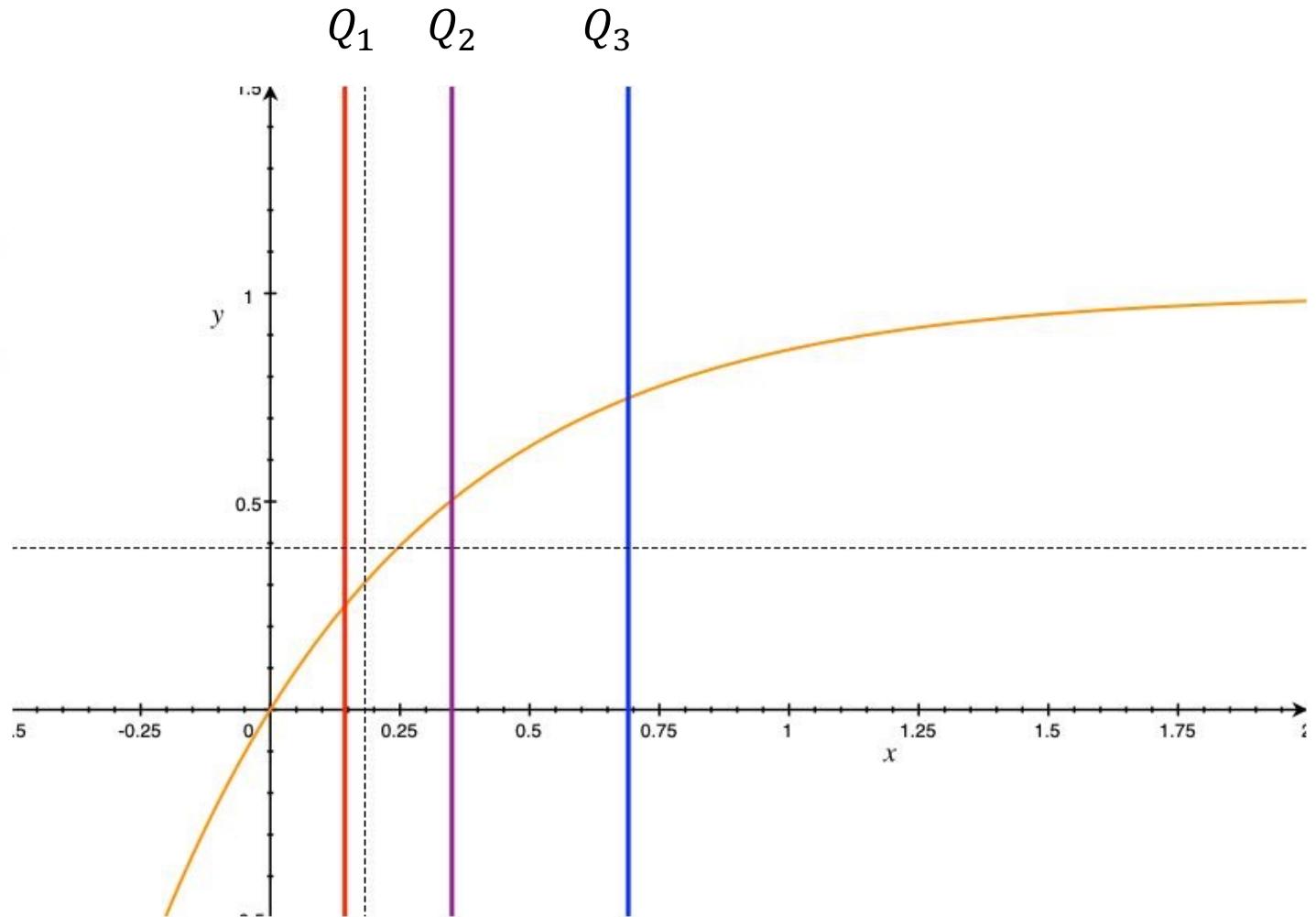
EXAMPLE

So we have

$$Q_1 = -\frac{\ln(1 - \frac{1}{4})}{2} \approx 0.1438$$

$$Q_2 = -\frac{\ln(1 - \frac{2}{4})}{2} \approx 0.3466$$

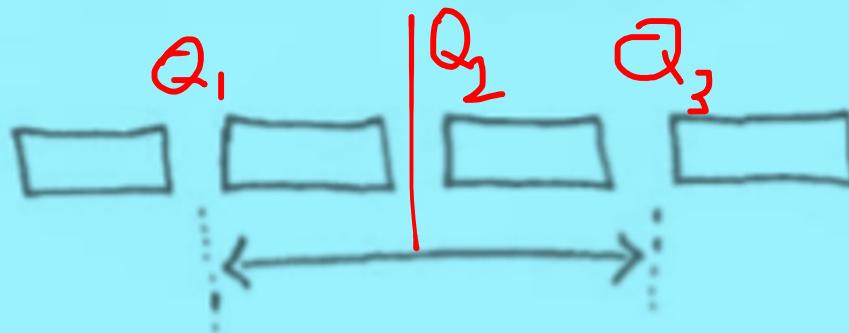
$$Q_3 = -\frac{\ln(1 - \frac{3}{4})}{2} \approx 0.6931$$



USE OF THE QUARTILES

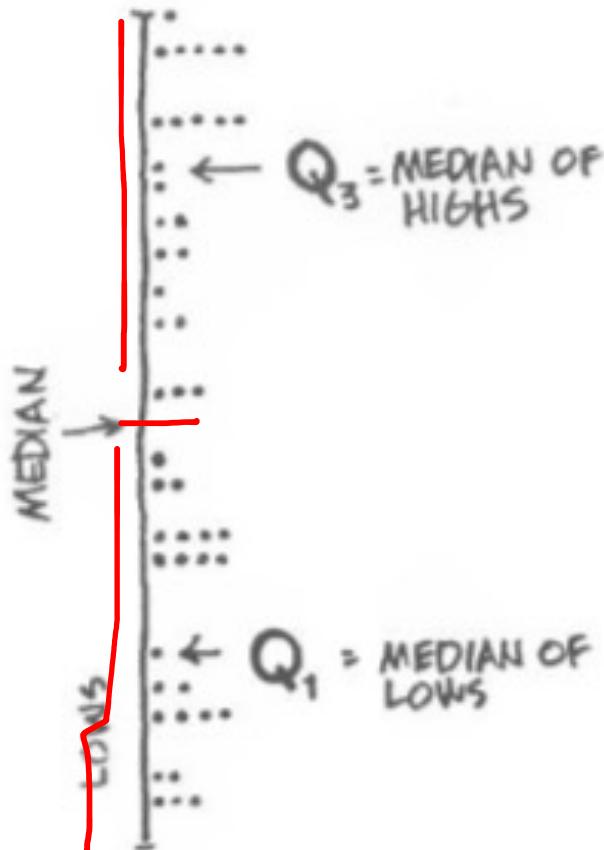
AGAIN, THERE'S MORE THAN ONE WAY TO MEASURE A SPREAD. ONE WAY IS
INTERQUARTILE RANGE

THE IDEA IS TO DIVIDE
THE DATA INTO FOUR
EQUAL GROUPS AND SEE
HOW FAR APART THE
EXTREME GROUPS ARE.



FIND QUARTILES IN A SET OF DATA

- 1) PUT THE DATA IN NUMERICAL ORDER.
- 2) DIVIDE THE DATA INTO TWO EQUAL HIGH AND LOW GROUPS AT THE MEDIAN. (IF THE MEDIAN IS A DATA POINT, INCLUDE IT IN BOTH THE HIGH AND LOW GROUPS.)
- 3) FIND THE MEDIAN OF THE LOW GROUP. THIS IS CALLED THE FIRST QUARTILE, OR Q_1 .
- 4) THE MEDIAN OF THE HIGH GROUP IS THE THIRD QUARTILE, OR Q_3 .

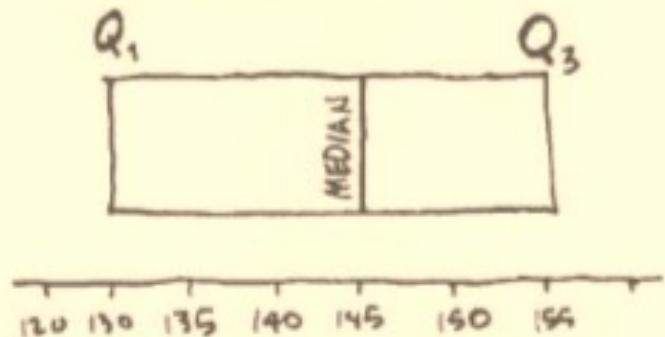


NOW THE INTERQUARTILE RANGE (IQR) IS THE DISTANCE (OR DIFFERENCE) BETWEEN THEM:

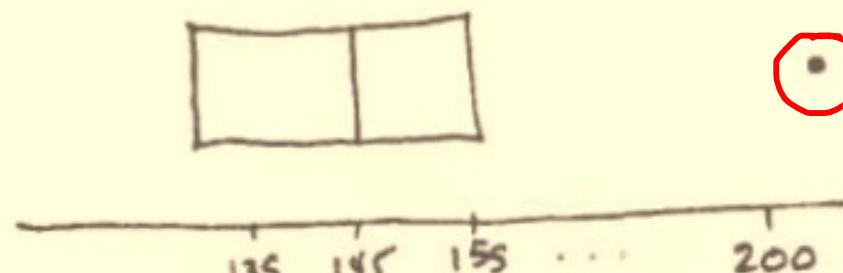
$$\text{IQR} = Q_3 - Q_1$$

BOX AND WHISKERS REPRESENTATION OF THE IQR

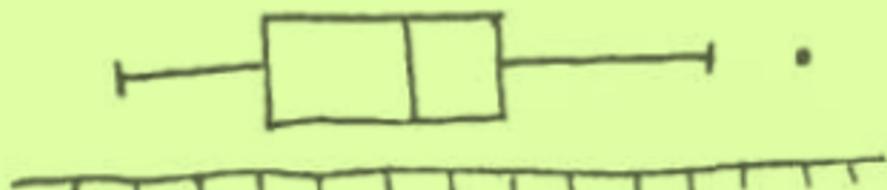
JOHN TUKEY INVENTED ANOTHER KIND OF DISPLAY TO SHOW OFF THE IQR, CALLED A BOX AND WHISKERS PLOT. THE BOX'S ENDS ARE THE QUARTILES Q_1 AND Q_3 . WE DRAW THE MEDIAN INSIDE THE BOX.



IF A POINT IS MORE THAN 1.5 IQR FROM AN END OF THE BOX, IT'S AN OUTLIER. DRAW THE OUTLIERS INDIVIDUALLY.



FINALLY, EXTEND "WHISKERS" OUT TO THE FARTHEST POINTS THAT ARE NOT OUTLIERS (I.E., WITHIN 1.5 IQR OF THE QUARTILES).



GRAPHICAL REPRESENTATION OF THE IQR



BOX-AND-
WHISKERS
PLOTS ARE
ESPECIALLY
GOOD FOR
SHOWING OFF
DIFFERENCES
BETWEEN
GROUPS.

PERCENTILES



Of course, there is no reason to restrict ourselves to quartiles.

We can also divide the interval in 100 parts and calculate
The so-called **percentiles**

We just need to set the cumulative distribution function equal to

$$F(x) = \frac{a}{100}$$

And chose the a^{th} percentile

Example

Consider a probability density function

$$f_Y(y) = \begin{cases} 2e^{-2y}, & y > 0, \\ 0, & \text{otherwise} \end{cases}$$

The cumulative distribution function is

$$F_Y(a) = \begin{cases} 0, & a < 0 \\ \int_0^a f_Y(y)dy = 1 - e^{-2a}, & a \geq 0 \end{cases}$$

Example

We can then calculate the n^{th} percentile, p_n , as

$$F_Y(p_n) = 1 - e^{-2p_n} = \frac{n}{100}$$

rearranging we get

$$p_n = -\frac{\ln(1 - \frac{n}{100})}{2}$$

So we have for instance

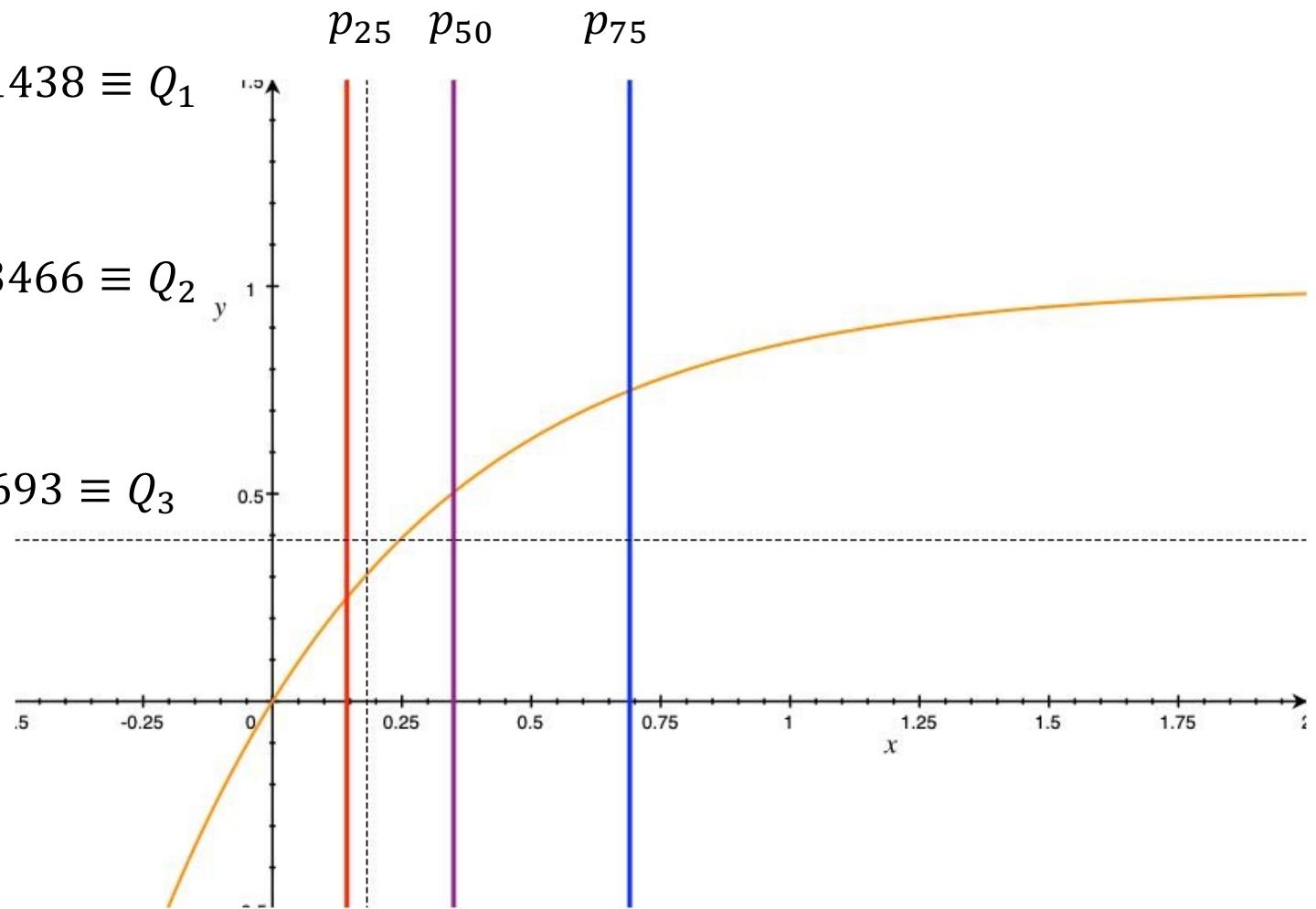
EXAMPLE

$$p_5 = -\frac{\ln\left(1 - \frac{5}{100}\right)}{2} \approx 0.026$$

$$p_{25} = -\frac{\ln\left(1 - \frac{25}{100}\right)}{2} \approx 0.1438 \equiv Q_1$$

$$p_{50} = -\frac{\ln\left(1 - \frac{50}{100}\right)}{2} \approx 0.3466 \equiv Q_2$$

$$p_{75} = -\frac{\ln\left(1 - \frac{75}{100}\right)}{2} \approx 0.693 \equiv Q_3$$



STATISTICAL TABLES



These give values of these percentiles for various distributions, that is all. Nowadays it is easier to get a computer to calculate them for you.

Particularly, in the IT sessions, we will use calls to the python **scipy library** to generate them.

SUMMARY: MEASURES OF VARIATION

We have already looked at how to calculate the variance and standard deviation of a random variable.

They measure how much the different values of a random variable differ from the mean of the variable, and how likely the given value is to occur.

- **Variance and standard deviation**
- **Range:** *largest value minus the smallest value.*
- **inter-quartile distance:** *the 'distance' between the first quartile and the third quartile*

These measures are often applied to samples, which may or may not come from some idealised probability distribution.

MARKOV'S INEQUALITY

Lets actually use all this machinery that we've built up to see if we can say anything general about a distribution.

Let X be a continuous random variable with probability density function $f(x)$.

We can show that if X only takes on positive values that

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

MARKOV'S INEQUALITY

Now

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \int_{-\infty}^a xf(x)dx + \int_a^{\infty} xf(x)dx \end{aligned}$$

Now we use the condition that X is non-negative to set up an inequality

$$\begin{aligned} E[X] &= \int_{-\infty}^a xf(x)dx + \int_a^{\infty} xf(x)dx \\ &\geq \int_a^{\infty} xf(x)dx \\ &\geq \int_a^{\infty} af(x)dx \\ &= a \int_a^{\infty} f(x)dx \\ &= aP(X \geq a) \end{aligned}$$

Which proves our initial statement.

CHEBYSHEV'S INEQUALITY

We've proved that

$$P(X \geq a) \leq \frac{E[X]}{a}$$

Take the particular case of $a = k^2$. And consider the expectation value of $(X - E[X])^2$

$$P((X - E[X])^2 \geq k^2) \leq \frac{E[(X - E[X])^2]}{k^2} = \frac{\sigma_X^2}{k^2}$$

As long as the mean of X , μ , and its variance, σ^2 , are finite then we can equivalently say

$$P(|X - E[X]| \geq k) \leq \frac{\sigma^2}{k^2}$$

CHEBYSHEV'S INEQUALITY

or if we define a new k' in units of σ_X , the standard deviation, with $k = k'\sigma_X$), we get

$$P(|X - \mathbb{E}[X]| \geq k'\sigma_X) \leq \frac{1}{k'^2}$$



CHEBYSHEV'S INEQUALITY



Pafnuty Lvovich Chebyshev (1821 –1894)
was a Russian mathematician.

His name can be alternatively transliterated as Chebychev, Chebysheff, Chebyshov; or Tchebychev, Tchebycheff (French transcriptions); or Tschebyschev, Tschebyschef, Tschebyscheff (German transcriptions)!

IMPLICATION OF CHEBYSHEV'S INEQUALITY

What does

$$P(|X - \mathbb{E}[X]| \geq k'\sigma_X) \leq \frac{1}{k'^2}$$

tell us?

Lets take $k' = 2$ to be concrete. Then

$$P(|X - \mathbb{E}[X]| \geq 2\sigma_X) \leq \frac{1}{4}$$

or put another way, 75% of the probability distribution is within 2 standard deviations of the mean.

OR, also that only at most $\frac{1}{3^2} = 0.111 \Rightarrow 11.1\%$ of the data will be further than $k'=3$ standard deviations from the mean.

NOTE: We have made no assumptions at all about the distribution - it applies to any random variable! The Chebyshev's inequality is generally a rather loose bound, if we know something about the real distribution we can do (much) better.

EXAMPLE



Suppose it is known that the number of items produced in a factory during a week is a random variable, X , with the mean value of 50.

What can be said about the probability that this week's production exceeds 75?

EXAMPLE

we can use Markov's inequality

$$P(X \geq a) \leq \frac{E[X]}{a}$$

here we know that $E[X]$ is 50, and the production is a random variable that does not take on negative values.

Plugging in the numbers we have

$$P(X \geq 75) \leq \frac{50}{75} = \frac{2}{3}$$

So not very useful in this case.

EXAMPLE

Q: If the production variance equals 25, then what can be said about the probability that the production this week will be between 40 and 60 items?

By Chebyshevs inequality in the form

$$P(|X - \text{E}[X]| \geq k) \leq \frac{\sigma^2}{k^2}$$

we can plug in the numbers to get

$$P(|X - 50| \geq 10) \leq \frac{25}{10^2} = \frac{1}{4}$$

so the probability that the production is between 40 and 60 is at least $1 - \frac{1}{4} = \frac{3}{4}$

EQUIVALENT PROBABILITIES



In the last expression we manipulated equivalent probabilities

$$\begin{aligned} P(|X - 50| \geq 10) &= P(X - 50 > 10 \cup X - 50 < -10) \\ &= P(X > 60 \cup X < 40) \\ &= P(40 < X < 60)^C \\ &= 1 - P(40 < X < 60) \end{aligned}$$

If we remember the things in the probabilities are really sets, we can rearrange them to get them into a useful form, and get different statements that define the same sets and hence the same probabilities.

WEAK LAW OF LARGE NUMBERS

We're getting a bit ahead of ourselves but let us go ahead and follow through a remarkable implication of the inequality we've just derived.

Let X_1, X_2, \dots, X_n be a sequence of independent and identically distributed random variables, each having $E[X] = \mu$. Then for any $\epsilon > 0$,

$$P\left(\left|\frac{X_1 + X_2 + \dots + X_n}{n} - \mu\right| > \epsilon\right) \rightarrow 0 \text{ as } n \rightarrow \infty$$

WEAK LAW OF LARGE NUMBERS: EXAMPLE

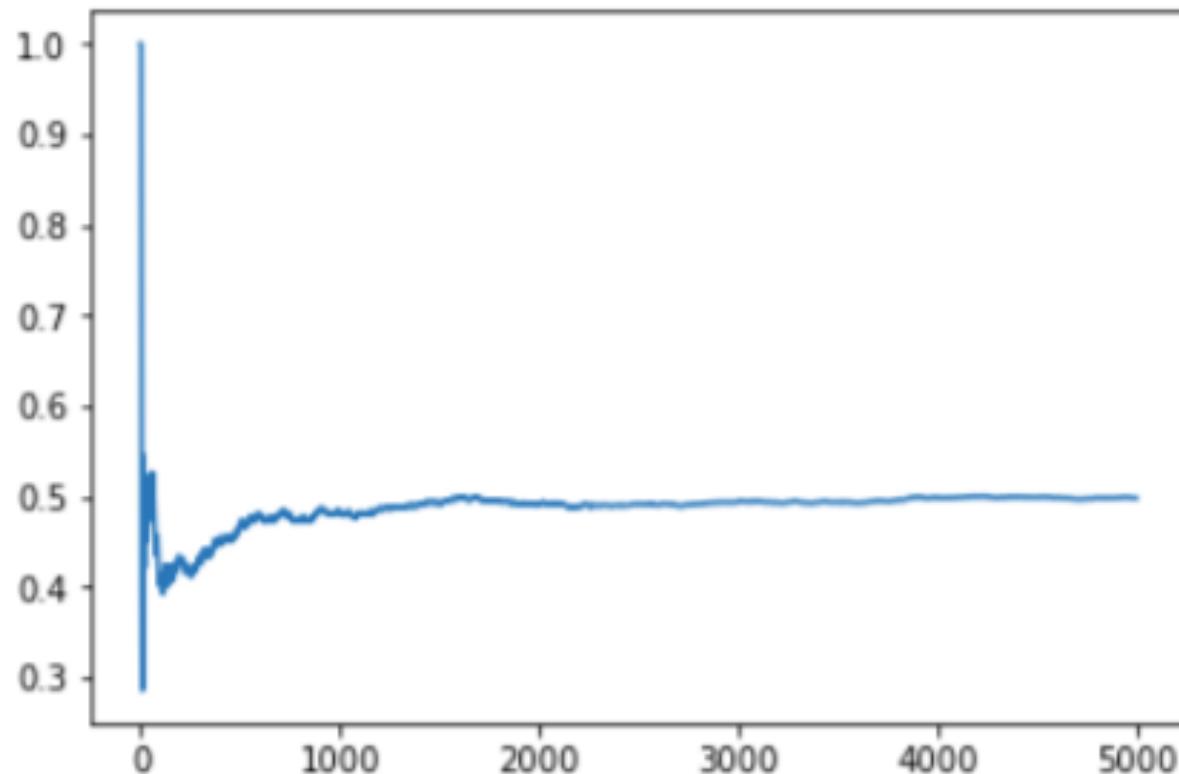
If the X_i correspond to consecutive tosses of a coin, and we give them a value 0 if they come up a tail and 1 a head then the weak law of large numbers will say that

$$P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \frac{1}{2}\right| > \epsilon\right) \rightarrow 0 \text{ as } n \rightarrow \infty$$

In this case, as $E[X_i] = P(\{\text{heads}\})$ we see that the average proportion of the first n flips that differ from the idealised probability $P(E)$ by more than $\epsilon \rightarrow 0$ as $N \rightarrow \infty$.

WEAK LAW OF LARGE NUMBERS: EXAMPLE

Expectation value of our variable, which equals $P(\{\text{heads}\})$, as we extend the length of a simulation.



SUMMARY



- Measures of Central tendency
- Medians, Quantiles and Percentiles
- Measures of Variation

as well as a remarkable result about probability distributions called

- Chebyshev's Inequality.

Which leads to

The Weak Law of Large Numbers



UNIVERSITY OF
LINCOLN

SCHOOL OF MATHEMATICS
AND PHYSICS

MTH1005

PROBABILITY AND STATISTICS

Semester B
Lecture 9 (9/4/2024)

Danilo Roccatano

Office:

INB 3323

Email:

droccatano@lincoln.ac.uk

SPECIAL DISTRIBUTIONS



Learning objectives:

Define and use the properties of the major probability distributions:

- **Uniform**
- Binomial
- Geometric
- Poisson
- Exponential
- Normal

with examples where these distributions occur.

UNIFORM RANDOM VARIABLE



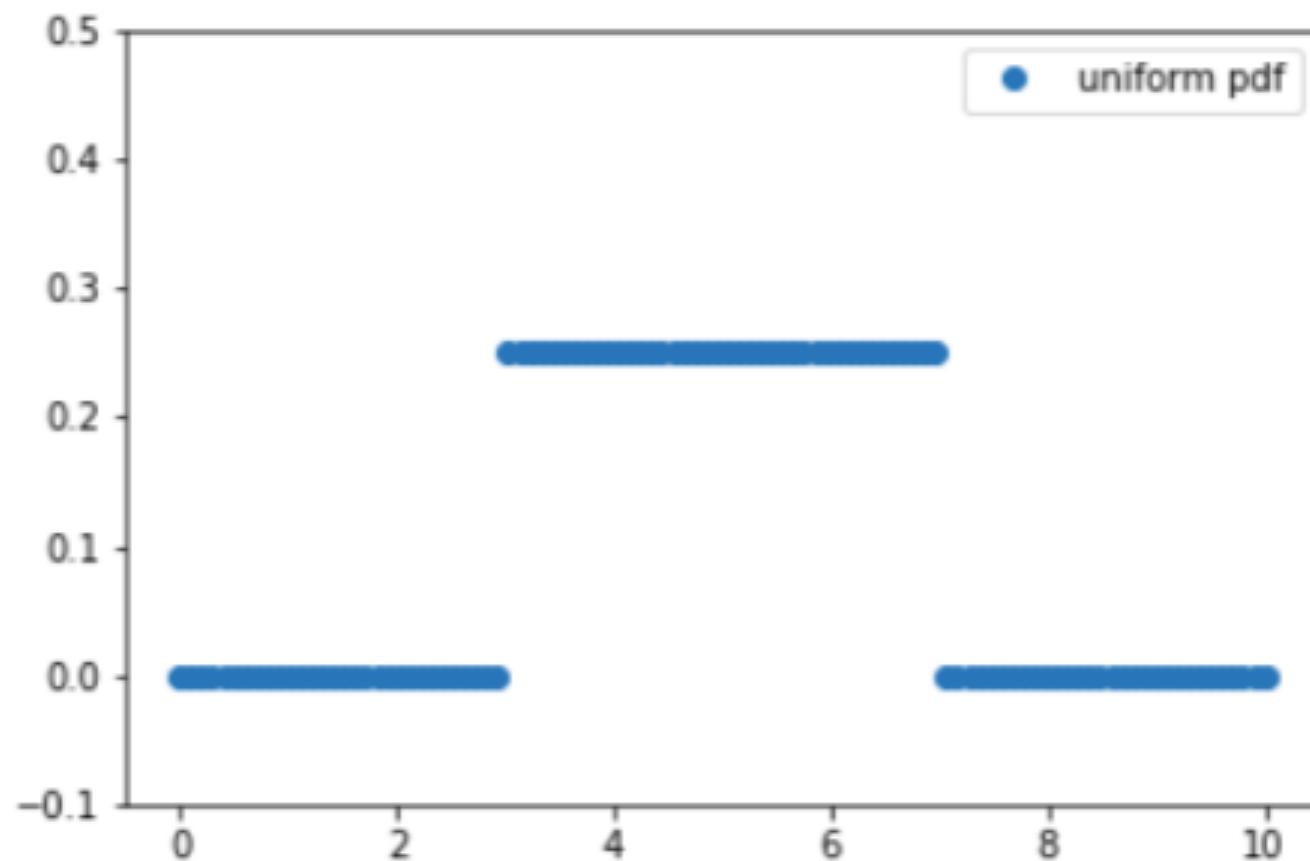
A continuous random variable X is said to be uniformly distributed over the interval $[a, b]$ if its probability density function is given

By 1

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \text{otherwise} \end{cases}$$

UNIFORM RANDOM VARIABLE: EXAMPLE

A uniform random variable between the values of 3 and 7 is shown below. It has a constant value of 14 over the interval [3,7].



DISCRETE UNIFORMLY DISTRIBUTED RANDOM VARIABLE



A discrete random variable Y would be described as uniformly distributed if all values of its probability mass function are equal over a finite domain.

We have used this fairly often in fact - it is used implicitly in all the '*counting problem*' types of situations: you assign probabilities by counting up the number of ways something can happen, then dividing by the total number of ways. This assumes a uniform distribution for the individual ways.

Examples

- Number of dots on a die
- students in the UoL School of Maths and Physics

PROPERTIES OF THE UNIFORM DISTRIBUTION

We can find the probability that a uniformly distributed random variable has a value in a subinterval $[\alpha, \beta]$ of $[a, b]$ using our normal definition:

$$\begin{aligned} P(\alpha \leq x \leq \beta) &= \int_{\alpha}^{\beta} f(x)dx \\ &= \frac{1}{(b-a)} \int_{\alpha}^{\beta} dx \\ &= \frac{(\beta - \alpha)}{(b-a)} \end{aligned}$$

this is just the ratio of the length of the subinterval $[\alpha, \beta]$ to that of the interval $[a, b]$ that the variable is defined over.

PROPERTIES OF THE UNIFORM DISTRIBUTION

This immediately shows that the integral of $f(x)$ over all space gives 1, as a probability density function must.

$$\begin{aligned} P(-\infty \leq x \leq \infty) &= \int_{-\infty}^{\infty} f(x)dx \\ &= \frac{1}{(b-a)} \int_a^b dx \\ &= \frac{(b-a)}{(b-a)} \\ &= 1 \end{aligned}$$

EXPECTATION VALUE

The expectation value of a uniformly distributed random variable is calculated as

$$\begin{aligned}\mathbb{E}[X] &= \int_a^b x f(x) dx \\&= \frac{1}{(b-a)} \int_a^b x dx \\&= \frac{b^2 - a^2}{2(b-a)} \\&= \frac{(b-a)(b+a)}{2(b-a)} \\&= \frac{a+b}{2}\end{aligned}$$

So the expected value of the random variable is the midpoint of the interval that it is defined over.

VARIANCE OF UNIFORMLY DISTRIBUTED RANDOM VARIABLE

We use our second definition of variance

$$\sigma_X^2 = E[X^2] - E[X]^2$$

We have already found $E[X]$, we need to calculate $E[X^2]$.

$$\begin{aligned} E[X^2] &= \int_a^b x^2 f(x) dx \\ &= \frac{1}{(b-a)} \int_a^b x^2 dx \\ &= \frac{b^3 - a^3}{3(b-a)} \\ &= \frac{b^2 + ab + a^2}{3} \end{aligned}$$

VARIANCE OF UNIFORMLY DISTRIBUTED RANDOM VARIABLE



and so we get

$$\begin{aligned}\sigma_X^2 &= E[X^2] - E[X]^2 \\ &= \frac{b^2 + ab + a^2}{3} - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{(b-a)^2}{12}\end{aligned}$$

QUANTILES

We can find the quantiles of the uniform distribution very simply.

The cumulative probability distribution function of a uniformly distributed continuous random variable defined between a and b is given by a

$$\begin{aligned} F(\alpha) &= \int_{-\infty}^a f(x)dx \\ &= \frac{1}{(b-a)} \int_a^\alpha dx \\ &= \frac{(\alpha - a)}{(b - a)} \end{aligned}$$

this is just the ratio of the length of the interval the variable is defined over to that of lower bound of the interval to the value of the variable at which we want to calculate the cumulative distribution function.

Example

Find the value of p_{95} such that there is a 95% probability that the random variable X takes a value smaller than p_{95} .

$$\begin{aligned} P(X < p_{95}) &= F(p_{95}) = 0.95 \\ &= \frac{(p_{0.95} - a)}{(b - a)} \\ \implies p_{0.95} &= 0.95(b - a) + a \end{aligned}$$

also implying the median is at the centre of the distribution, like the mean.

RANDOM NUMBER



Probably, the most important use of the uniform distribution is in defining a random number.

The value of a uniform $[0,1]$ random variable is called a random number.

We used computer approximations to this distribution in our simulations.

In Scientific Computing module, you will learn more about the generation of random numbers using computers.

SPECIAL DISTRIBUTIONS



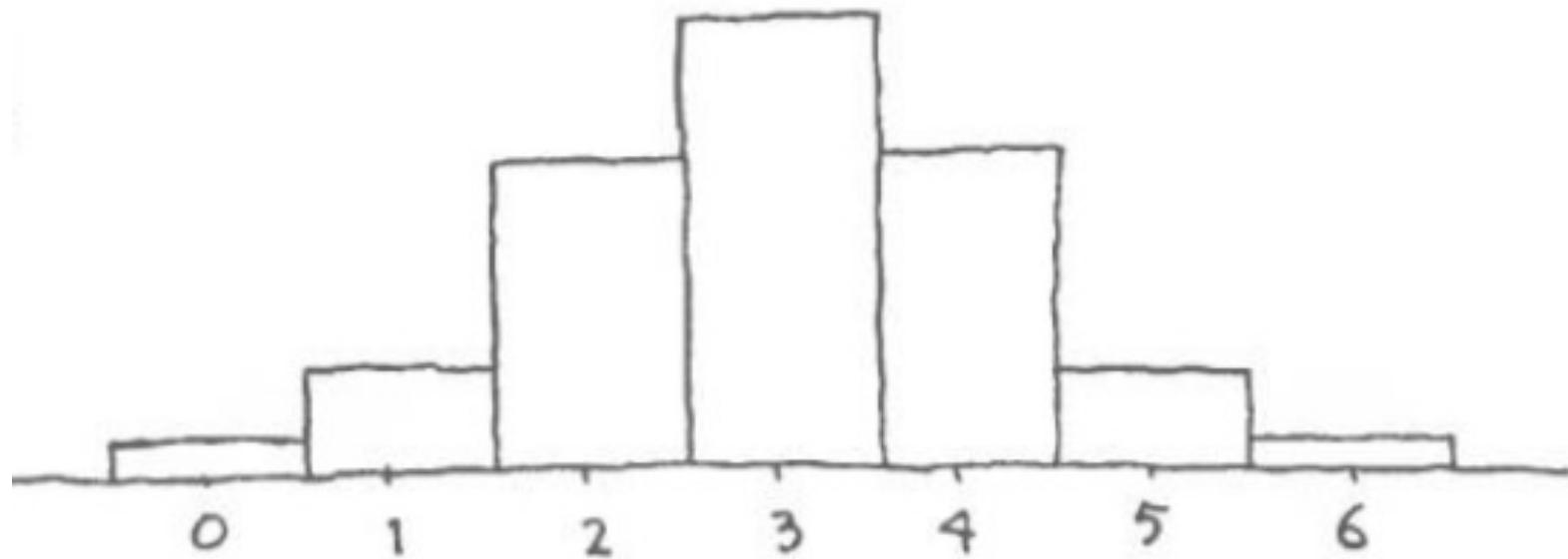
Learning objectives:

Define and use the properties of the major probability distributions:

- Uniform
- **Binomial**
- Poisson
- Geometric
- Exponential
- Normal

with examples where these distributions occur.

BINOMIAL DISTRIBUTION



BINOMIAL DISTRIBUTION

We've used this fairly often in the first few weeks.

Definition

It is the discrete distribution that arises when we have an experiment that can be partitioned into two mutually exclusive events ($\{S\}$, success and $\{F\}$, failure) with a probability of success $P(\{S\})$ that is known, repeated independently n times.

Definition

A discrete random variable X with a probability mass function of the form:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, (x = 0, 1, \dots, n - 1, n)$$

is said to have a binomial distribution.

Where n is the number of repetitions of the experiment and p is the probability of 'success'.

CHECK IT IS A PROBABILITY MASS DISTRIBUTION

To be a probability mass distribution we must have

$$\sum_{\text{all } x} P(x) = 1$$

For the binomial distribution we have

$$\begin{aligned}\sum_{\text{all } x} P(x) &= \sum_{\text{all } x} \binom{n}{x} p^x (1-p)^{n-x} \\ &= (p + (1-p))^n \quad \text{we use the binomial expansion} \\ &= 1^n = 1 \qquad \qquad \qquad (a + b)^n = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k\end{aligned}$$

So it fits that criteria, and all $p(x)$ are between 0 and 1.

THE MEAN AND THE VARIANCE

We have that the mean (expectation value) of our binomially distributed variable X is

$$\begin{aligned} E[X] &= \sum_{\text{all } x} xp(x) \\ &= \sum_{\text{all } x} x \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=1}^n x \frac{n! p^x (1-p)^{n-x}}{(n-x)! x!} \quad (x = 0 \text{ term is zero}) \\ &= np \sum_{x=1}^n x \frac{(n-1)! p^{x-1} (1-p)^{n-x}}{(n-x)! x!} \quad (\text{factorize out } np) \end{aligned}$$

THE MEAN AND THE VARIANCE

- $\frac{x}{x!} = \frac{1}{(x-1)!}$
- $(n - x)!$ can be rewritten as $(n - 1 - (x - 1))!$
- $(1 - p)^{n-x}$ can be written as $(1 - p)^{n-1-(x-1)}$ we then get

$$E[X] = np \sum_{x=1}^n \frac{(n-1)!p^{x-1}(1-p)^{n-1-(x-1)}}{(n-1-(x-1))!(x-1)!}$$

and, we substitute $y = x - 1$ and $N = n - 1$ to get

$$E[X] = np \sum_{y=0}^N \frac{N!p^y(1-p)^{N-y}}{(N-y)!y!}$$

$$\begin{aligned} &= np \left[\sum_{y=0}^N \binom{N}{y} p^y (1-p)^{N-y} \right] \\ &= np \end{aligned}$$

THE MEAN AND THE VARIANCE



The variance can be found in a similar manner, but I won't give it all here.

$$\begin{aligned}\sigma_X^2 &= E[X^2] - E[X]^2 \\ &= n^2 p^2 - np^2 + np - n^2 p^2 \\ &= np(1 - p)\end{aligned}$$

Note: the same proof can be obtained in a more elegant way using moment generating functions.

SUMMARY OF EXPECTATION AND VARIANCE OF BINOMIAL DISTRIBUTION



EXPECTATION

$$E[X] = np$$

VARIANCE

$$\sigma_X^2 = E[X^2] - E[X]^2 = np(1 - p)$$

CUMULATIVE DISTRIBUTION FUNCTION



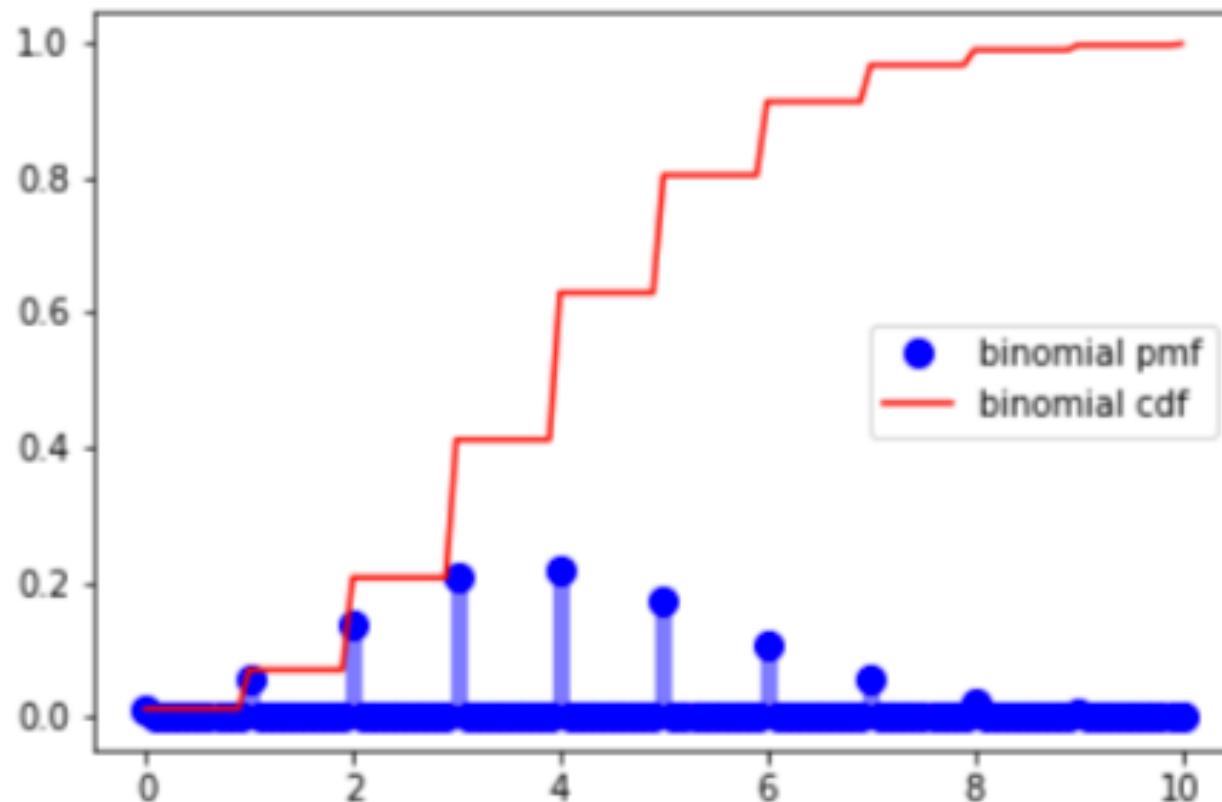
For discrete random variables, the cumulative distribution function is often available in tabulated form.

To calculate it we need to sum up all the $p(x)$ values where x is less than the value of interest.

$$F(y) = \sum_{x \leq y} p(x)$$

this can get quite tricky for the binomial distribution.

CUMULATIVE DISTRIBUTION FUNCTION



CUMULATIVE DISTRIBUTION FUNCTION



we can access the quantiles for the binomial distribution like so

```
stats.binom.ppf(0.99, n, p)  
6.0
```

This says that for our distribution (binomial with parameters n and p) a random value taken from this distribution will have a value less than than 6.99% of the time.

SPECIAL DISTRIBUTIONS



Learning objectives:

Define and use the properties of the major probability distributions:

- Uniform
- Binomial
- **Poisson**
- Geometric
- Exponential
- Normal

THE POISSON DISTRIBUTION

Recherches sur la probabilité des jugements en matière criminelle et en matière civile (1837).

The distribution was introduced to describe the theoretical number of wrongful convictions in a given country by focusing on certain random variables N that count, among other things, the number of discrete occurrences (sometimes called "events" or "arrivals") that take place during a time-interval of given length



Siméon Denis Poisson (1781–1840)
Image source:wikipedia

THE POISSON DISTRIBUTION



Soon after its introduction, the distribution was used for practical applications. For example,

In 1860, Simon Newcomb fitted the Poisson distribution to the number of stars found in a unit of space.

In 1898 Ladislaus Bortkiewicz, used the distribution to investigate the number of soldiers in the Prussian army killed accidentally by horse kicks.

In 1907, William Sealy Gosset used it to calculate the number of yeast cells used when brewing Guinness beer.

1909, A.K. Erlang, used it for the number of phone calls arriving at a call centre within a minute.

In 1946, D. Clarke used it to investigate the targeting of V-1 flying bombs on London during World War II.

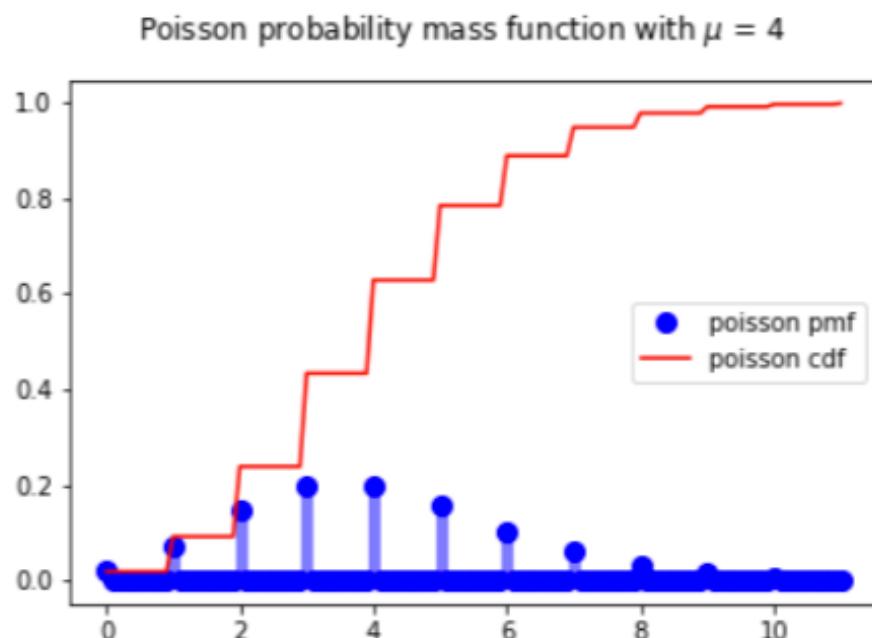
POISSON DISTRIBUTION

Definition: a random variable X having a probability mass function of the form:

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

and μ can take on any positive value, is said to have a Poisson distribution.

What does it look like?



POISSON DISTRIBUTION



It is a discrete probability distribution function that describes the number of times an event that occurs infrequently in a given time period will occur.

Other examples are:

- misprints on a page in a book
- number of wrong telephone numbers dialled in a day
- number of α -particles discharged in a fixed period of time from a radioactive nuclei.
- the counts of prime numbers in short intervals (Gallagher, 1976).

CHECK IT IS A VALID PROBABILITY MASS FUNCTION

We have that

$$\sum_{\text{all } x} p(x) = \sum_{x=0}^{\infty} e^{-\mu} \frac{\mu^x}{x!}$$

$$= e^{-\mu} \sum_{x=0}^{\infty} \frac{\mu^x}{x!}$$

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

$$= e^{-\mu} e^{\mu}$$

$$= 1$$

as required.

We use the series definition of the exponential function in the next to last step.

We've not dealt with discrete variables with infinite numbers of outcomes as far as we are concerned, they are OK as long as all properties and probabilities that we need to calculate converge as $x \rightarrow \infty$.

EXPECTATION AND VARIANCE OF A POISSON DISTRIBUTED RANDOM VARIABLE



If X has $P(x) = \frac{e^{-\mu} \mu^x}{x!}$

then

- $E[X] = \mu$
- $\sigma^2 = \mu$

PROOF OF EXPECTATION OF POISSON DISTRIBUTED RANDOM VARIABLE

$$\begin{aligned} E[X] &= \sum_{x=0}^{\infty} xe^{-\mu} \frac{\mu^x}{x!} \\ &= \sum_{x=1}^{\infty} xe^{-\mu} \frac{\mu^x}{x!} \\ &= e^{-\mu} \sum_{x=1}^{\infty} \frac{\mu^x}{(x-1)!} \\ &= \mu e^{-\mu} \sum_{x=1}^{\infty} \frac{\mu^{(x-1)}}{(x-1)!} \end{aligned}$$

then we substitute $y = x - 1$

$$\begin{aligned} &= \mu e^{-\mu} \sum_{y=0}^{\infty} \frac{\mu^y}{(y)!} \\ &= \mu e^{-\mu} e^{\mu} \\ &= \mu \end{aligned}$$

PROOF OF VARIANCE OF POISSON DISTRIBUTED RANDOM VARIABLE

$$\begin{aligned}\mathbb{E}(X^2) &= \sum_{x=0}^{\infty} x^2 e^{-\mu} \frac{\mu^x}{x!} \\&= \sum_{x=1}^{\infty} (x(x-1) + x) e^{-\mu} \frac{\mu^x}{x!} \\&= \sum_{x=2}^{\infty} x(x-1) e^{-\mu} \frac{\mu^x}{x!} + \sum_{x=1}^{\infty} x e^{-\mu} \frac{\mu^x}{x!} \\&= \mu^2 e^{-\mu} \sum_{x=2}^{\infty} \frac{\mu^{(x-2)}}{(x-2)!} + \mathbb{E}[X] \\&= \mu^2 + \mu\end{aligned}$$

where we substitute in $y = x - 2$, similarly to the expectation value.

PROOF OF VARIANCE OF POISSON DISTRIBUTED RANDOM VARIABLE



Then identify the sum as an exponential function.

We then get that

$$\begin{aligned}\sigma_X^2 &= E[X^2] - E[X]^2 \\ &= \mu^2 + \mu - \mu^2 \\ &= \mu\end{aligned}$$

WHEN IS IT USED?



If an event with a low probability of occurrence at any instant is

1. known *to occur randomly, and*
2. *to have a mean number of occurrences μ in a given time (or space) interval*

has a Poisson distribution (approximately) with parameter μ .

An example



Government statistic for a certain country show that the average number of major industrial accidents per year for large firms is 1.1 per 5000 employees. Find the probability that there will be at least one major accident per year for firm with:

- 5000 employees
- 10000 employees

SOLUTION

Here we have a unit scale of 5000 employees and we are given the average accident rate per year is 1.1 per 5000 employees.

This should satisfy our criteria to be Poisson distributed - the accidents are rare and randomness should be a reasonable assumption.

Then we have a Poisson distribution with $\mu = 1.1$

$$P(X \geq 1) = 1 - p(0) = 1 - e^{-1.1} \frac{1.1^0}{0!} = 0.6671$$

in the interval of interest (1 year).

SOLUTION



Now we have a new scale of 10000 employees. If the accidents are rare, independent and random we'd then expect 2.2 accidents per year in a firm with 10000 employees. So

$$P(X \geq 1) = 1 - p(0) = 1 - e^{-2.2} \frac{2.2^0}{0!} = 0.8892$$

APPROXIMATION TO BINOMIAL DISTRIBUTION



One way to see how it can occur is to see how it can approximate a binomial distribution if n is large and p small.

Recall that

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, (x = 0, 1, \dots, n - 1, n)$$

for a binomially distributed random variable.

APPROXIMATION TO BINOMIAL DISTRIBUTION

So if X is a binomially distributed random variable with parameters (n, p) . If we let $\mu = np$

$$\begin{aligned} P(X = x) &= \binom{n}{x} p^x (1 - p)^{n-x} \\ &= \frac{n!}{(n - x)! x!} p^x (1 - p)^{n-x} \\ &= \frac{n!}{(n - x)! x!} \frac{\mu^x}{n^x} \left(1 - \frac{\mu}{n}\right)^{n-x} \\ &= \frac{n(n - 1) \cdots (n - x + 1)}{n^x} \frac{\mu^x}{x!} \frac{\left(1 - \frac{\mu}{n}\right)^n}{\left(1 - \frac{\mu}{n}\right)^x} \end{aligned}$$

APPROXIMATION TO BINOMIAL DISTRIBUTION

Now for large n and small p we can make the approximations

$$\begin{aligned} \left(1 - \frac{\mu}{n}\right)^n &\approx e^{-\mu}, \\ \frac{n(n-1)\cdots(n-x+1)}{n^x} &\approx 1, \\ \left(1 - \frac{\mu}{n}\right)^x &\approx 1 \end{aligned}$$

Then we get that

$$P(X = x) \approx e^{-\mu} \frac{\mu^x}{x!}$$

SUMMARY OF DISTRIBUTION FUNCTIONS

If X has

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, (x = 0, 1, \dots, n - 1, n)$$

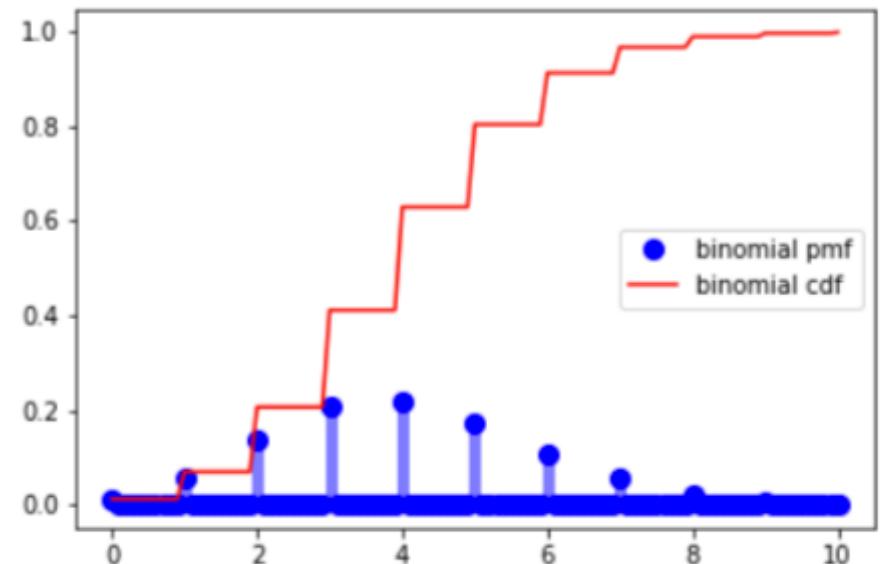
Then it is a **binomial distribution** with

EXPECTATION

$$\mathbb{E}[X] = np$$

VARIANCE

$$\sigma_X^2 = E[X^2] - E[X]^2 = np(1 - p)$$



SUMMARY OF DISTRIBUTION FUNCTIONS

If X has

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

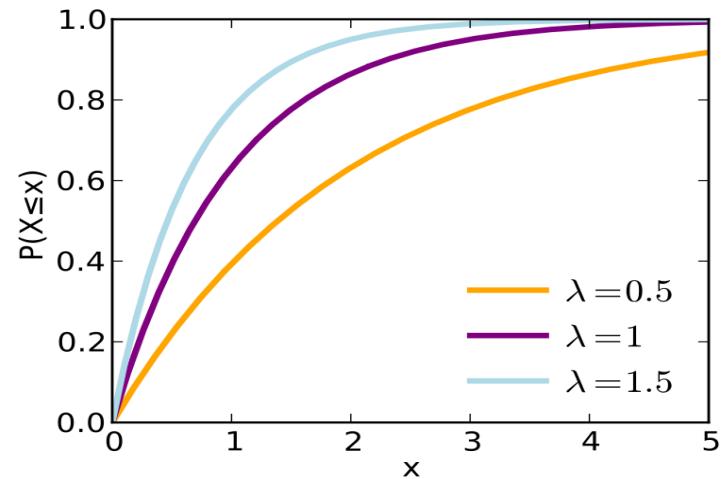
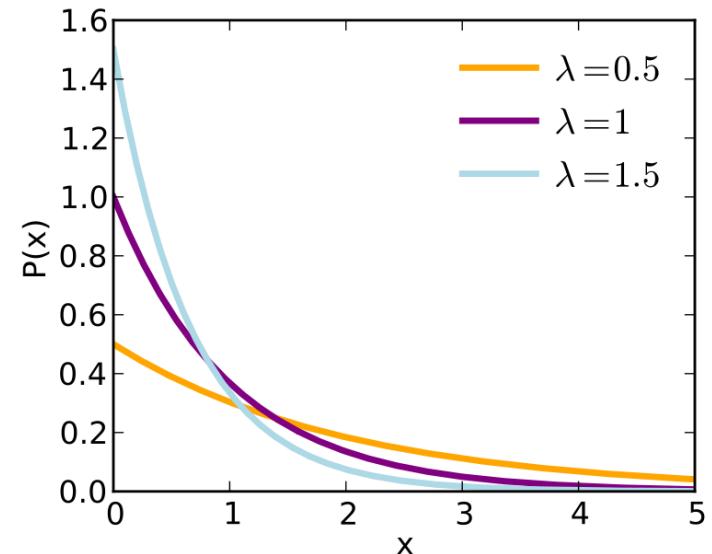
then it is a **Poisson distribution** with

EXPECTATION

$$\mathbb{E}[X] = \mu$$

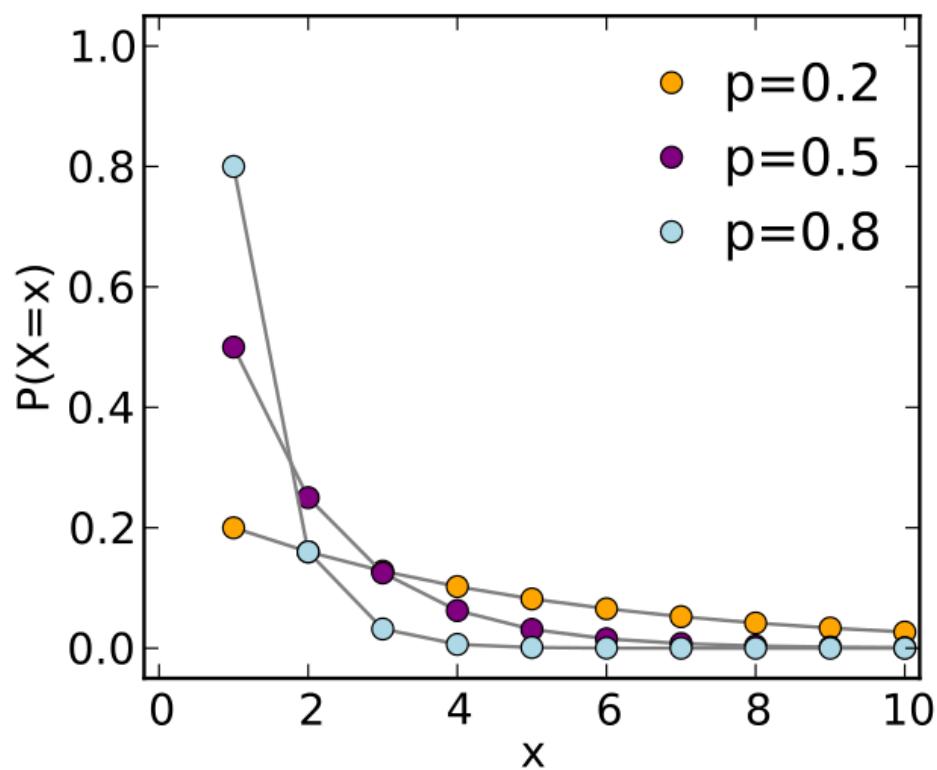
VARIANCE

$$\sigma^2 = \mu$$



Cumulative distribution
function

GEOMETRIC DISTRIBUTION



GEOMETRIC DISTRIBUTION



A **discrete random variable** X with probability mass function $p(x)$ of the form:

$$P(X = x) = p(x) = (1 - p)^{x-1} p$$

with $0 \leq p \leq 1$ is said to have a **geometric distribution**.

You can prove that it is a valid probability mass function by noting that the sum over all x is a geometric progression that can be summed up immediately.

GEOMETRIC DISTRIBUTION

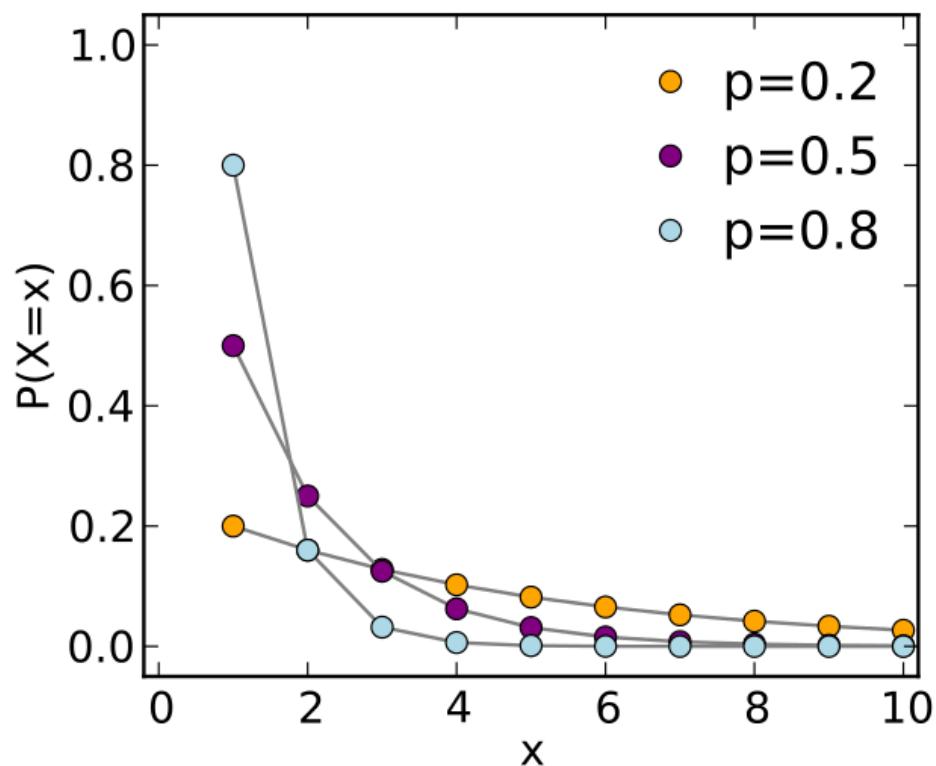


$$P(X = x) = p(x) = (1 - p)^{x-1} p$$

For $q=(1-p)$ and $k=x-1$, we can write that

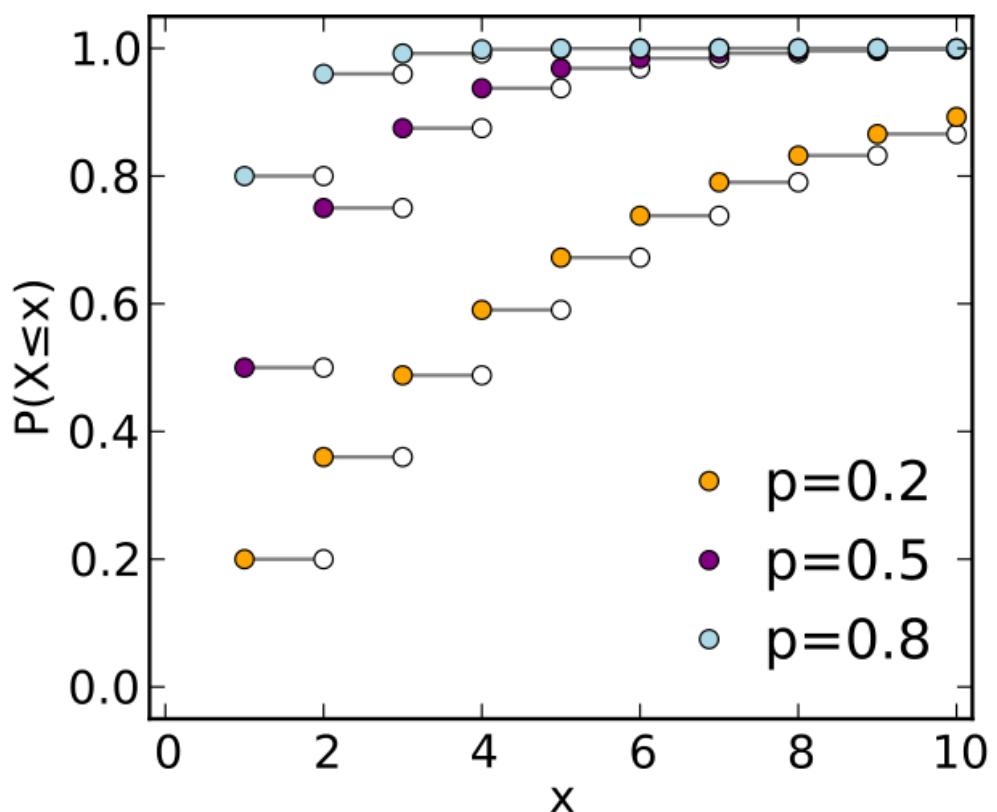
$$\sum_{x=1}^{\infty} p(1 - p)^{x-1} = \sum_{x=1}^{\infty} pq^k = \frac{p}{1 - q} = \frac{p}{1 - 1 + p} = 1$$

GEOMETRIC DISTRIBUTION



Probability
distribution
function

Cumulative
distribution
function



Source: wikipedia

GEOMETRIC DISTRIBUTION

We just give the expectation and variance here for completeness.

$$E[X] = \frac{1-p}{p}$$

$$\text{var}(X) = \frac{1-p}{p^2}$$

Note that the way in which this distribution is used can give different results:

- Some define the variable as the number of experiments until we obtain a success. In this case, we get $E[X] = \frac{1}{p}$.
- Others talk about the expected number of failures before a success. In this case, we get $E[x] = \frac{1-p}{p}$.

EXAMPLE

What is the expected number of rolls of a fair die to obtain a 1?

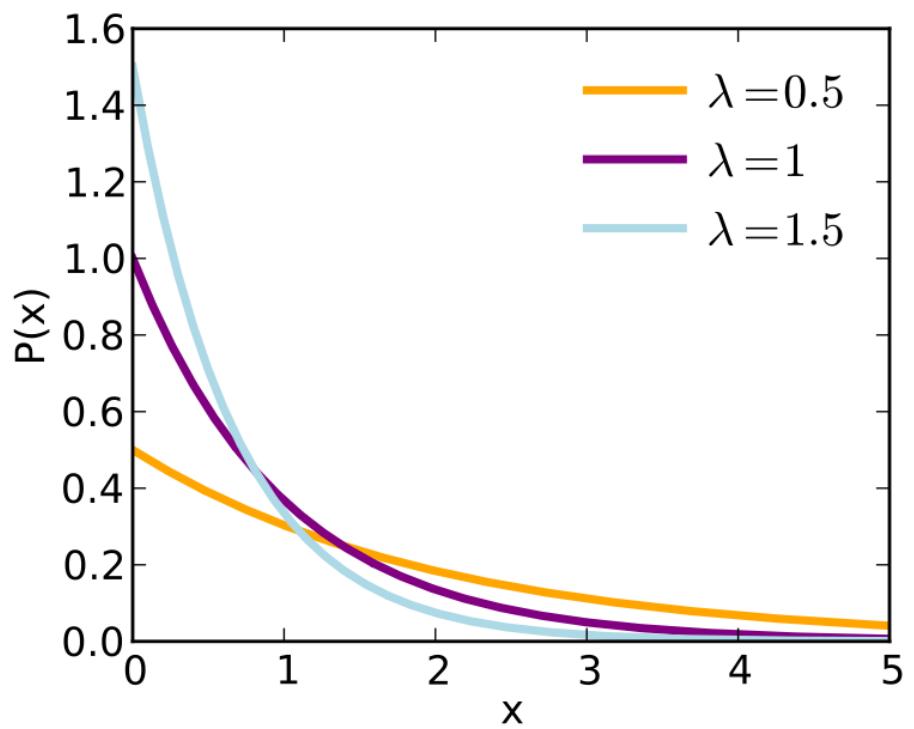
Here we formulate the question as the total number of experiments until we get a success. So, we have

$$E[X] = \frac{1}{\frac{1}{6}} = 6$$

If it had been phrased as the number of rolls before we get the 1 we'd have

$$E[X] = \frac{1 - \frac{1}{6}}{\frac{1}{6}} = 5$$

EXPONENTIAL DISTRIBUTION



EXPONENTIAL DISTRIBUTION



The exponential distribution function of a continuous random variable y is the following

$$\begin{aligned}f_Y(y) &= \beta e^{-\beta y}, y > 0 \\&= 0 \text{ otherwise:}\end{aligned}$$

Lets find the expectation and variance using the moment generating function of Y .

EXPONENTIAL DISTRIBUTION

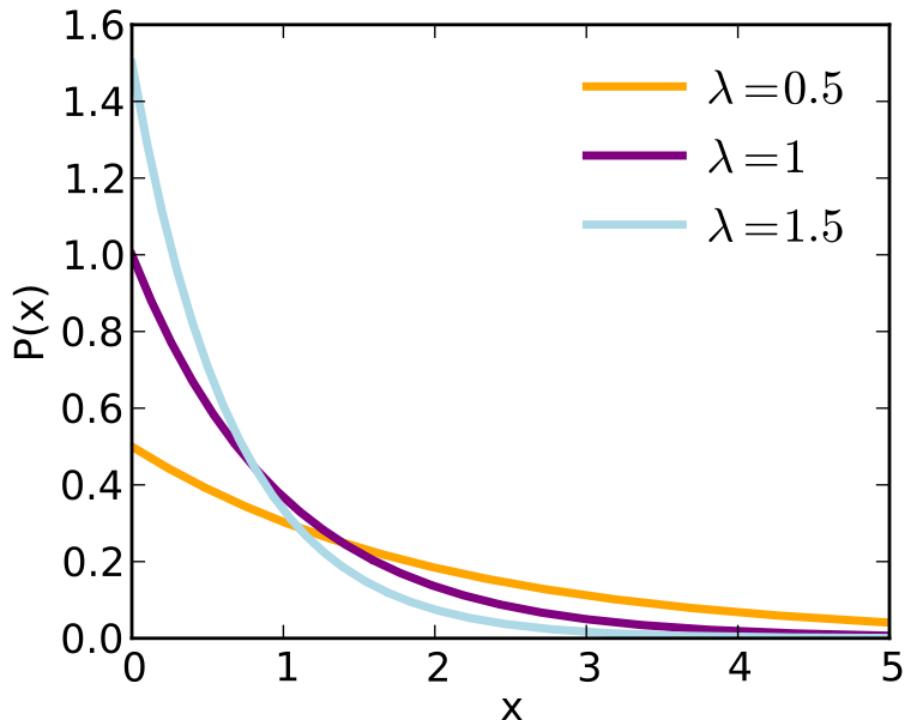


The exponential distribution occurs naturally when describing the lengths of the inter-arrival times in a homogeneous Poisson process.

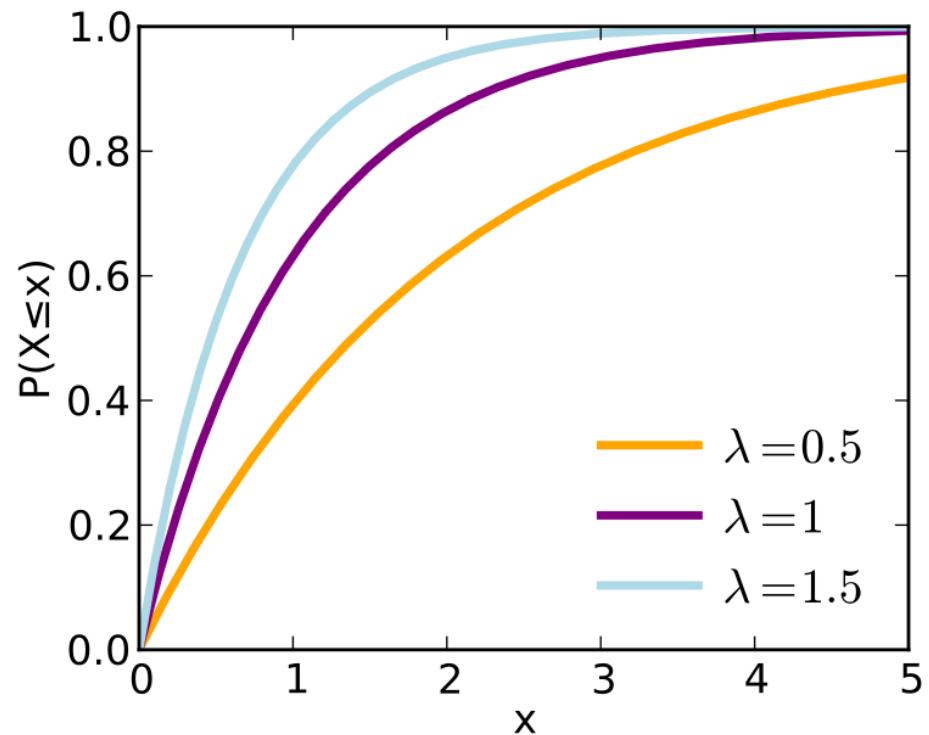
For example:

- The time until a radioactive particle decays, or the time between clicks of a Geiger counter
- The time it takes before your next telephone call
- How long it takes for a bank teller etc. to serve a customer.
- Distance between DNA mutation in a gene.

EXPONENTIAL DISTRIBUTION



Probability distribution
function



Cumulative distribution
function

MOMENT GENERATING FUNCTION

The moment-generating function (MGF) of a random variable X is a tool used in probability theory and statistics to conveniently calculate moments of the distribution associated with Moments provide important summary statistics of a probability distribution, capturing information about its central tendency, spread, and shape.

The moment generating function, denoted by $m_X(t)$, is defined as the expected value (or mean) of e^{tX} , where t is a parameter.

Mathematically, it is expressed as:

$$m_X(t) = E[e^{tX}], \quad -\infty < t < \infty$$

where E denotes the expected value operator.

MOMENT GENERATING FUNCTION

Using the MGF, we can compute moments of the distribution of X by taking derivatives of $m_X(t)$ with respect to t and then evaluating the result at $t=0$.

Recall the Taylor series of e^{tx} about $x=0$:

$$e^{tx} = 1 + tx + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \dots$$

MOMENT GENERATING FUNCTION

By substituting in the MGF we obtain

$$\begin{aligned}m_X(t) &= \mathbb{E}[e^{tX}] \\&= \mathbb{E}\left[1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \dots\right]\end{aligned}$$

and again, using the linear properties of $\mathbb{E}[\cdot]$ we get

$$\begin{aligned}m_X(t) &= 1 + \mathbb{E}[tX] + \frac{t^2}{2!}\mathbb{E}[X^2] + \frac{t^3}{3!}\mathbb{E}[X^3] + \dots \\&= 1 + tm_1 + \frac{t^2}{2!}m_2 + \frac{t^3}{3!}m_3 + \dots\end{aligned}$$

MOMENT GENERATING FUNCTION

If we now take derivatives with respect to t we get

$$\frac{dm_X(t)}{dt} = m_1 + tm_2 + \frac{t^2}{2!}m_3 + \dots$$
$$\frac{d^2m_X(t)}{dt^2} = m_2 + tm_3 + \frac{t^2}{2!}m_4 + \dots$$

Finally setting $t=0$ all terms in the series will be equal to zero except the first

$$\left. \frac{dm_X(t)}{dt} \right|_{t=0} = m_1$$
$$\left. \frac{d^2m_X(t)}{dt^2} \right|_{t=0} = m_2$$

Etc.

EXPONENTIAL DISTRIBUTION

The moment generating function of Y is

$$\begin{aligned}m_Y(t) &= E[e^{tY}] \\&= \int_0^\infty e^{ty} \beta e^{-\beta y} dy \\&= \int_0^\infty \beta e^{-y(\beta-t)} dy \\&= \frac{\beta}{\beta - t}\end{aligned}$$

$$m^{(1)}(0) = \frac{d(m_t(0))}{dt} = \frac{\beta}{(\beta - t)^2} = \beta^{-1}$$

to find the moments we take the derivative with respect to t then set $t=0$

$$\begin{aligned}\mu_Y &= m^{(1)}(0) = \beta^{-1} \\m_2 &= m^{(2)}(0) = 2\beta^{-2}\end{aligned}$$

so

$$\sigma_Y^2 = m_2 - \mu^2 = \beta^{-2}$$

EXAMPLE

The length of lifetime a of transistor is a random variable y with density function

$$f_{Y(y)} = \begin{cases} 0.001e^{-0.001y} & y > 0 \\ 0 & Otherwise \end{cases}$$

the moment generating function of Y is

$$\begin{aligned} m_Y(t) &= E[e^{tY}] \\ &= \int_0^{\infty} e^{ty} 0.001e^{-0.001y} dy \\ &= \int_0^{\infty} 0.001e^{-y(0.001-t)} dy \\ &= \frac{0.001}{0.001 - t} \end{aligned}$$

EXAMPLE

to find the moments we take the derivative with respect to t then set t

$$\begin{aligned}\mu &= m^{(1)}(0) = 1000 \\ m_2 &= m^{(2)}(0) = 2(1000)^2\end{aligned}$$

$$\begin{aligned}m^1 &= \frac{0.001}{(0.001 - t)^2} \\ m^2 &= \frac{2(0.001)}{(0.001 - t)^3}\end{aligned}$$

so

$$\sigma^2 = m_2 - \mu^2 = (1000)^2$$

SUMMARY

If X has

$$f_Y(y) = \beta e^{-\beta y}, y > 0 \\ = 0 \text{ otherwise:}$$

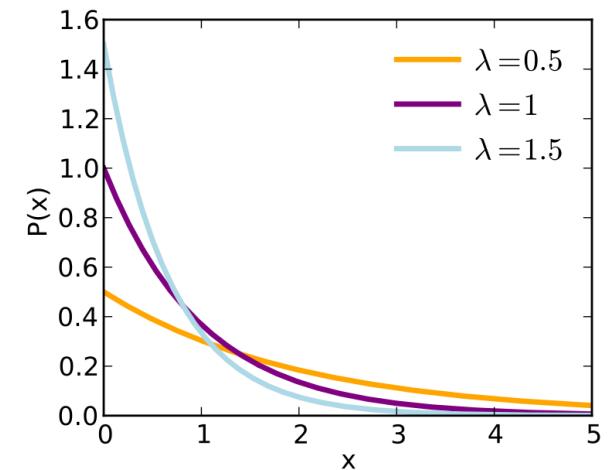
Then it is a **exponential distribution** with

EXPECTATION

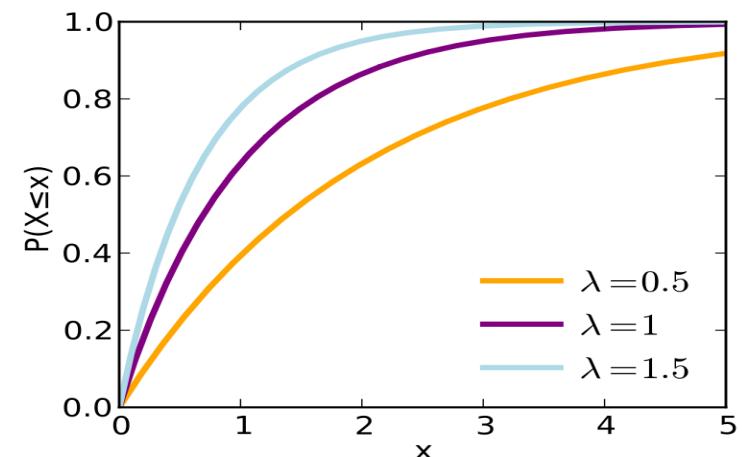
$$\mu_Y = m^{(1)}(0) = \beta^{-1}$$

VARIANCE

$$\sigma_Y^2 = m_2 - \mu^2 = \beta^{-2}$$



PMF



Cumulative distribution function