Week 0~~2~~3 notes

Repeated trials

01 Theory

Repeated trials

When a single experiment type is repeated many times, and we assume each instance is *independent* of the others, we say it is a sequence of **repeated trials** or **independent trials**.

The probability of any sequence of outcomes is derived using independence together with the probabilities of outcomes of each trial.

A simple type of trial, called a **Bernoulli trial**, has two possible outcomes, 1 and 0, or success and failure, or T and F. A sequence of repeated Bernoulli trials is called a **Bernoulli process**.

- Write sequences like $TFFTTF$ for the outcomes of repeated trials of this type.
- Independence implies

$$P[TFFTTF] = P[T] \cdot P[F] \cdot P[F] \cdot P[T] \cdot P[T] \cdot P[F]$$

- Write $p = P[T]$ and $q = P[F]$, and because these are all outcomes (exclusive and exhaustive), we have $q = 1 - p$. Then:

$$P[TFFTTF] \quad \gg\gg \quad pqqppq \quad \gg\gg \quad p^3q^3 = \left(\tfrac{1}{2}\right)^6$$

- This gives a formula for the probability of any sequence of these trials.

$\frac{1}{2^6} = \left(\frac{1}{2}\right)^6$ $\quad p = \frac{1}{2}$

$q = \frac{1}{2}$

A more complex trial may have three outcomes, A, B, and C.

- Write sequences like $ABBACABCA$ for the outcomes.
- Label $p = P[A]$ and $q = P[B]$ and $r = P[C]$. We must have $p + q + r = 1$.
- Independence implies

$$P[ABBACABCA] \quad \gg\gg \quad pqqprpqrp \quad \gg\gg \quad p^4q^3r^2$$

- This gives a formula for the probability of any sequence of these trials.

Let S stand for the *sum of successes* in some Bernoulli process. So, for example, "$S = 3$" stands for the event that the number of successes is exactly 3. The probabilities of S events follow a **binomial distribution**.

Suppose a coin is biased with $P[H] = 20\%$, and H is 'success'. Flip the coin 20 times. Then:

$$P[S = 3] \quad \gg\gg \quad \binom{20}{3} \cdot 0.2^3 \cdot 0.8^{17}$$

Each outcome with exactly 3 heads and 17 tails has probability $0.2^3 \cdot 0.8^{17}$. The *number* of such outcomes is the number of ways to choose 3 of the flips to be heads out of the 20 total flips.

The probability of at least 18 heads would then be:

$$P[S \geq 18] \quad \gg\gg \quad P[S = 18] + P[S = 19] + P[S = 20]$$

$$\gg\gg \quad \binom{20}{18} \cdot (0.2)^{18} \cdot (0.8)^2 + \binom{20}{19} \cdot (0.2)^{19} \cdot (0.8)^1 + \binom{20}{20} \cdot (0.2)^{20} \cdot (0.8)^0$$

With three possible outcomes, A, B, and C, we can write sum variables like S_A which counts the number of A outcomes, and S_B and S_C similarly. The probabilities of events like "$(S_A, S_B, S_C) = (2, 3, 5)$" follow a **multinomial distribution**.

02 Illustration

Example - Multinomial: Soft drinks preferred

Folks coming to a party prefer Coke (55%), Pepsi (25%), or Dew (20%). If 20 people order drinks in sequence, what is the probability that exactly 12 have Coke and 5 have Pepsi and 3 have Dew?

Solution

The multinomial coefficient $\binom{20}{12, 5, 3}$ gives the number of ways to assign 20 people into bins according to preferences matching the given numbers, $C = 12$ and $P = 5$ and $D = 3$.

Each such assignment is one sequence of outcomes. All such sequences have probability $(0.55)^{12} \cdot (0.25)^5 \cdot (0.2)^3$.

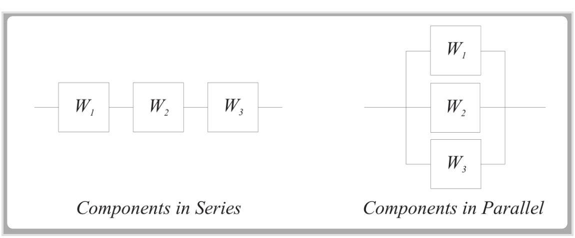
The answer is therefore:

$$\binom{20}{12, 5, 3} \cdot (0.55)^{12} \cdot (0.25)^5 \cdot (0.2)^3 \quad \gg\gg \quad \frac{20!}{12!\ 5!\ 3!} \cdot (0.55)^{12} \cdot (0.25)^5 \cdot (0.2)^3$$

Reliability

03 Theory

Consider some process schematically with **components in series** and **components in parallel**:



- Each component has a probability of success or failure.
- Event W_i indicates 'success' of that component (same name).
- Then $P[W_i]$ is the probability of W_i succeeding.

Success for a *series* of components requires success of *each* member.

- Series components *rely on each other*.
- Success of the whole is success of part 1 AND success of part 2 AND part 3, etc.

Failure for *parallel* components requires failure of *each* member.

- Parallel components represent *redundancy*.
- Success of the whole is success of part 1 OR success of part 2 OR part 3, etc.

For series components:

$$P[W] = P[W_1W_2W_3] = P[W_1] \cdot P[W_2] \cdot P[W_3]$$

For parallel components:

$$P[W^c] = \text{"failure"} \quad \gg\gg \quad P[W_1^cW_2^cW_3^c]$$

$$\gg\gg \quad (1 - P[W_1])(1 - P[W_2])(1 - P[W_3])$$

If $P[W_i] = p$ for all components i, then:

- Series components: $P[W] = p^3$
- Parallel components: $P[W] = 1 - (1 - p)^3$

To analyze a complex diagram of series and parallel components, bundle each:

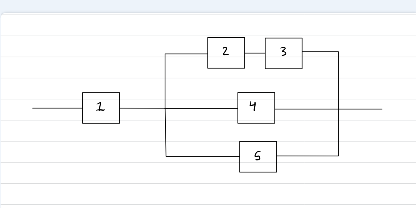
- pure series set as a single compound component with its own success probability (the product)
- pure parallel set as a single compound component with its own success probability (using the failure formula)

This is like the analysis of resistors and inductors.

04 Illustration

Example - Series, parallel, series

Suppose a process has internal components arranged like this:



Write W_i for the event that component i succeeds, and W_i^c for the event that it fails.

The success probabilities for each component are given in the chart:

1	2	3	4	5
92%	89%	95%	86%	91%

Find the probability that the entire system succeeds.

Solution

1. Conjoin components 2 and 3 in series.
   - Compute:

$$P[W_2W_3] \quad \gg\gg \quad P[W_2]\cdot P[W_3] \quad \gg\gg \quad (0.89)\cdot(0.95) = 0.8455$$

   - Therefore:

$$P\big[(W_2W_3)^c\big] \quad \gg\gg \quad 1-0.846 \quad \gg\gg \quad 0.1545$$

2. Conjoin components (2-3) with 4 and 5 in parallel.
   - Compute for the complement (failure) first:

$$P\big[(W_2W_3 \cup W_4 \cup W_5)^c\big] \quad \gg\gg \quad P[(W_2W_3)^c]\cdot P[W_4^c]\cdot P[W_5^c]$$

$$\gg\gg \quad (0.1545)(0.14)(0.09) \quad \gg\gg \quad 0.0019467$$

   - Flip back to success:

$$P\big[W_2W_3 \cup W_4 \cup W_5\big] \quad \gg\gg \quad 1-0.0019467 \quad \gg\gg \quad 0.9980533$$

3. Conjoin components 1 with (2-3-4-5) in series.
   - Compute:

$$P\Big[W_1\big(W_2W_3 \cup W_4 \cup W_5\big)\Big] \quad \gg\gg \quad (0.92)(0.9980533)$$

$$\gg\gg \quad 0.918209036 \quad \approx 91.82\%$$

Discrete random variables

05 Theory

Random variable

A **random variable (RV)** X on a probability space $(S,\mathcal{F},P)$ is a function $X:S\to\mathbb{R}$.

So X assigns to each *outcome* a *number*.

- The word ‘variable’ indicates that the RV outputs *numbers*.

Random variables can be formed from other random variables using mathematical operations on the output numbers.

Given random variables X and Y, we can form these new ones:

$$\frac{1}{2}(X+Y), \qquad X\cdot Y, \qquad \cos X, \qquad X^2, \qquad \text{etc.}$$

Suppose $s\in S$ is some particular outcome. Then, for example, $(X+Y)(s)$ is by definition $X(s)+Y(s)$.

Random variables determine events.

- Given $a\in\mathbb{R}$, the event “$X=a$” is equal to the set $X^{-1}(a)$
- That is: the set of outcomes mapped to a by X
- That is: the event “X took the value a”

Such events have probabilities. We write them like this:

$$P[X = a] \quad \gg\gg \quad P[X^{-1}(\overset{a}{\cancel{s}})]$$

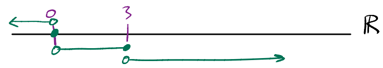
This generalized to events where X lies in some range or set, for example:

$$P[a \leq X < b], \qquad P\big[X \in \{2, 4, 5, 6, 9\}\big]$$

$P[X^{-1}(\{2,4,5,6,9\})]$

The axioms of probability translate into rules for these events.

For example, additivity leads to:



$$P[X < 0] + P[X = 0] + P[0 < X \leq 3] + P[3 < X] = 1$$

A **discrete** random variable has probability concentrated at a discrete set of real numbers.

- A 'discrete set' means *finite or countably infinite*.
- The distribution of probability is recorded using a **probability mass function (PMF)** that assigns probabilities to each of the discrete real numbers.
- Numbers with nonzero probability are called **possible values**.

$\Phi(-Z) = 0.84$

$1 - \Phi(+Z) = 0.84$

$\Phi(+Z) = 0.16$

$\Phi(0.16) =: Z$

$\Phi^{-1}(0.84) = \text{"}a\text{"}$

$Z = -a$

PMF

The PMF function for X (a discrete RV) is defined by:

$$P_X(k) := P[X = k]$$

for $k \in \mathbb{R}$ a possible value.

A **continuous** random variable has probability spread out over the space of real numbers.

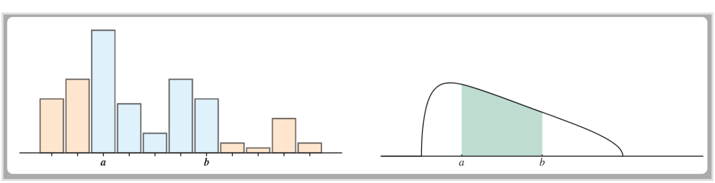
- The distribution of probability is recorded using a **probability density function (PDF)** which is *integrated over intervals* to determine probabilities.

PDF

The PDF function for Y (a CRV) is written $f_Y(x)$ for $x \in \mathbb{R}$, and probabilities are calculated like this:

$\int_{-\infty}^{\infty} f_Y(x)dx = 1$

$$P\big[a \leq \underset{Y}{\cancel{x}} \leq b\big] = \int_a^b f_{\underset{Y}{\cancel{X}}}(x)\,dx$$



For any RV, whether discrete or continuous, the distribution of probability is encoded by a function:

CDF

The **cumulative distribution function (CDF)** for a random variable X is defined for all $x \in \mathbb{R}$ by:

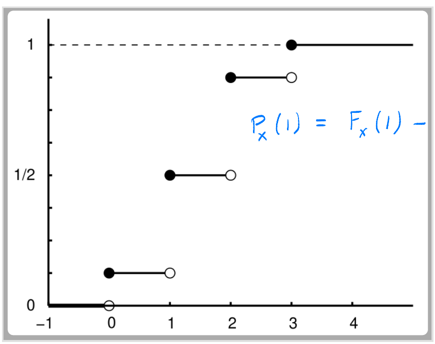
$$F_X(x) = P[X \leq x] \xrightarrow{\text{continuous}} F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$$

Notes:

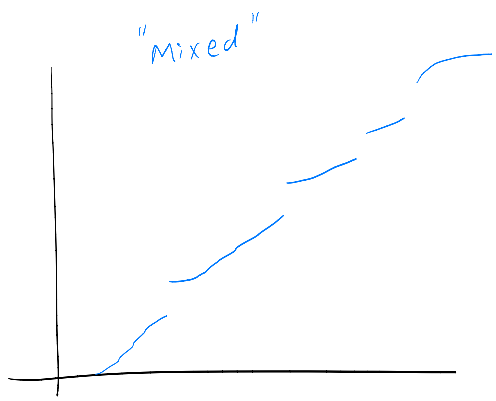
- Sometimes the relation to X is omitted and one sees just "$F(x)$."
- Sometimes the CDF is called, simply, "the distribution function" because:
- The CDF works equally well for discrete and continuous RVs.
    - Not true for PMF (discrete only) or PDF (continuous only).
    - There are *mixed* cases (partly discrete, partly continuous) for which the CDF is *essential*.

The CDF of a discrete RV is always a stepwise increasing function. At each step up, the jump size matches the PMF value there.

From this graph of $F_X(x)$:



$P_X(1) = F_X(1) - \lim_{t \to 1^-} F_X(t)$



we can infer the PMF values based on the jump sizes:

$P_X(-1)$	$P_X(0)$	$P_X(1)$	$P_X(2)$	$P_X(3)$	$P_X(4)$
0	1/8	3/8	3/8	1/8	0

For a discrete RV, the CDF and the PMF can be calculated from each other using formulas.

PMF from CDF from PMF

Given a PMF $P_X(x)$, the CDF is determined by:

$$F_X(x) = \sum_{k_i \leq x} P_X(k_i)$$

where $\{k_1, k_2, \ldots\}$ is the set of possible values of X.

Given a CDF $F_X(x)$, the PMF is determined by:

$$P_X(k) = F_X(k) - \lim_{x \to k^-} F_X(x) \quad = \text{"jump" at } k$$

06 Illustration

☰ Example - PDF and CDF: Roll 2 dice

Roll two dice colored red and green. Let X_R record the number of dots showing on the red die, X_G the number on the green die, and let S be a random variable giving the total number of dots showing after the roll, namely $S = X_R + X_G$.

- Find the PMFs of X_R and of X_G and of S.
- Find the CDF of S.
- Find $P[S = 8]$.

Solution

1. ☰ Sample space.
    - Denote outcomes with ordered pairs of numbers (i, j), where i is the number showing on the red die and j is the number on the green one.
    - Require that $i, j \in \mathbb{N}$ are integers satisfying $1 \leq i,\ j \leq 6$.
    - Events are sets of distinct such pairs.
2. Create chart of outcomes.
    - Chart:

$X_R \longrightarrow$

$X_G \downarrow$

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

3. ☰ Definitions of X_R, X_G, and S.
    - We have $X_R(i, j) = i$ and $X_G(i, j) = j$.
    - Therefore $S(i, j) = i + j$.
4. Find PMF of X_R.
    - Use variable n for each possible value of X_R, so $n = 1, 2, \ldots, 6$.
    - Find $P_{X_R}(n)$:

$$P_{X_R}(n) \quad \gg\gg \quad P[X_R = n]$$

$$\gg\gg \quad \frac{|\text{outcomes with } n \text{ on red}|}{|\text{all outcomes}|} \quad \gg\gg \quad \frac{6}{36} = \frac{1}{6}$$

    - Therefore $P_{X_R}(n) = 1/6$ for every n.
5. ☰ Find PMF of X_G.
    - Same as for X_R:

$$P_{X_G}(n) = \frac{1}{6} \quad \text{for all } n, \quad (=0 \text{ all else})$$

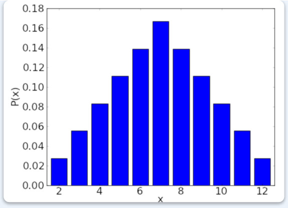
6. Find PMF of S.
    - Find $P_S(n)$:

$$P_S(n) \quad \gg\gg \quad P[S = n] \quad \gg\gg \quad \frac{|\text{outcomes with sum } n|}{|\text{all outcomes}|}$$

    - ⚠ Count outcomes along *diagonal lines* in the chart.
    - Create table of $P_S(n)$:

k	2	3	4	5	6	7	8	9	10	11	12
$p_S(k) = P(S = k)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

    - Create bar chart of $P_S(n)$:



- Evaluate: $P[S=8]$ $\quad \gg\gg \quad 5/36.$

7. Find CDF of S.
   - CDF definition:

$$F_S(n) = P[S \le n]$$

   - Apply definition: add new PMF value at each increment:

$$F_S(n) = \begin{cases} 0 & n < 1 \\ 1/36 & 1 \le n < 2 \\ 3/36 & 2 \le n < 3 \\ 6/36 & 3 \le n < 4 \\ 10/36 & 4 \le n < 5 \\ 15/36 & 5 \le n < 6 \\ 21/36 & 6 \le n < 7 \\ 26/36 & 7 \le n < 8 \\ 30/36 & 8 \le n < 9 \\ 33/36 & 9 \le n < 10 \\ 35/36 & 10 \le n < 11 \\ 36/36 & 11 \le n \end{cases}$$

+1 to all

☰ Example - Total heads count; binomial expansion of 1

A fair coin is flipped n times.

Let X be the random variable that counts the total number of heads in each sequence.

The PMF of X is given by:

P["X = k"] = P["k heads"]

$$P_X(k) = \binom{n}{k}\left(\frac{1}{2}\right)^n$$

Since the total probability must add to 1, we know this formula must hold:

$$1 = \sum_{\text{possible } k} P_X(k)$$

$$\gg\gg \quad 1 = \sum_{k=0}^{n} \binom{n}{k}\left(\frac{1}{2}\right)^n$$

Is this equation really true?

There is another way to view this equation: it is the binomial expansion $(x+y)^n$ where $x = \frac{1}{2}$ and $y = \frac{1}{2}$:

$$\left(\frac{1}{2}+\frac{1}{2}\right)^n = \sum_{k=0}^{n} \binom{n}{k}\left(\frac{1}{2}\right)^n$$

☰ Example - Life insurance payouts

A life insurance company has two clients, A and B, each with a policy that pays \$100,000 upon death. Consider events D_1 that the older client dies next year, and D_2 that the younger dies next year. Suppose $P[D_1] = 0.10$ and $P[D_2] = 0.05$.

Define a random variable X measuring the total money paid out next year in units of \$1,000. The possible values for X are 0, 100, 200. We calculate:

$$\begin{aligned} P[X = 0] \quad &\gg\gg \quad P[D_1^c]P[D_2^c] = 0.95 \cdot 0.90 = 0.86 \\ P[X = 100] \quad &\gg\gg \quad 0.05 \cdot 0.90 + 0.95 \cdot 0.10 = 0.14 \\ P[X = 200] \quad &\gg\gg \quad 0.05 \cdot 0.10 = 0.005 \end{aligned}$$