Probability & Entropy

Semester 1, 2021

Kris Ehinger

Outline

- Variable types
- Probability basics
- Probability distributions
- Entropy
- Bayes' rule

Variable types

Attribute types

- Each instance can have many attributes
- Attributes can be various types
  - Nominal (or categorical)
  - Ordinal
  - Continuous (or numerical)

Nominal/categorical variable

- Variable can take multiple values which are discrete types or categories
  - outlook: sunny, overcast, rainy
  - object: car, bike, pedestrian, street sign, …
- There is no natural ordering of the values; they are all equally dissimilar from each other
- **boolean** is a special type of nominal variable with only two possible values
  - isMonday: true, false

Ordinal variable

- Variable has discrete values, and they have a natural order
  - beverageSize: small medium large
  - rating: 
- Ordered but not real numeric values – mathematical relations don’t make sense, distances might not be consistent, can’t add/subtract them
- Thresholds are meaningful (e.g., >3 stars)
- Nominal/ordinal distinction can be unclear

Continuous/numerical variable

- Variable is real-valued with a defined zero point and no explicit bound
  - distance
  - time
  - price
- Intervals are consistent, mathematical operations make sense (e.g., 2m + 3m = 5m distance)
- What about `int` variables?
  - They take discrete values in the computer, but usually represent a continuous variable in the world

has mathematical relation

Attribute types

- Why does it matter?
  - Different variable types imply different types of structure and need to be handled differently in learning
  - Some models can only work with nominal or continuous data

Review: Variable types

- A researcher is modelling how changes in course fees would affect the number of students enrolling in various courses.
- What types are the variables in this model?
  - Course name (e.g., Law, Nursing, Engineering)
  - Course fee
  - Number of students

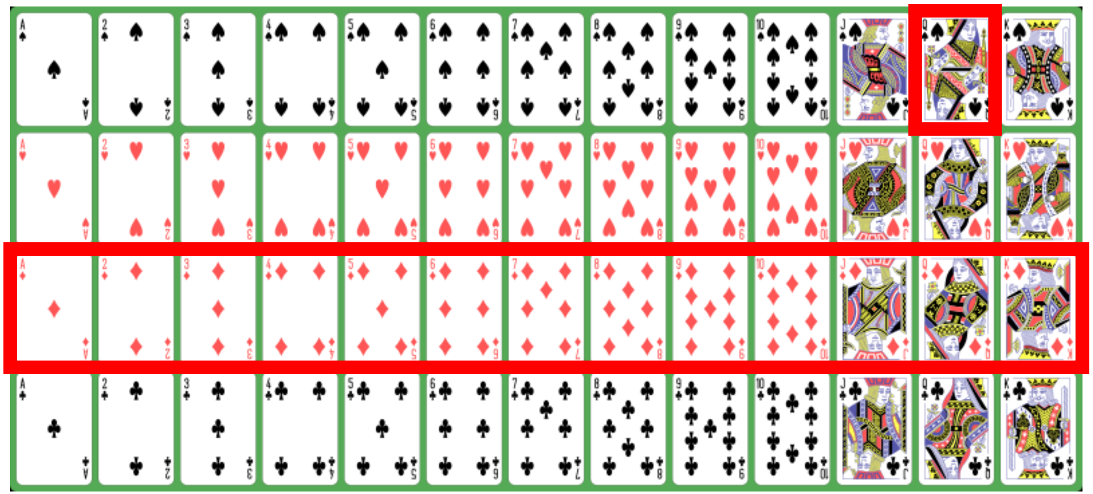
Probability basics

Why probability?

- Learning implies uncertainty
  - Problem too complex to solve exactly
  - Data is ambiguous or incomplete
- How to make smart decisions under uncertainty?

Probability notation

- P(x) = probability of an event x

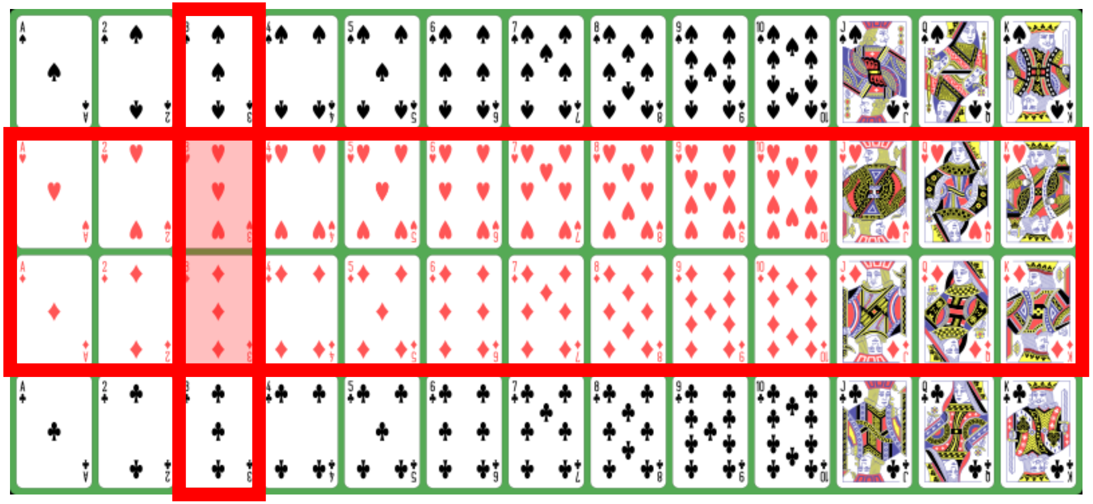


P(Queen of spades) = 1/52

P(diamond) = 13/52 = 1/4

Joint probability

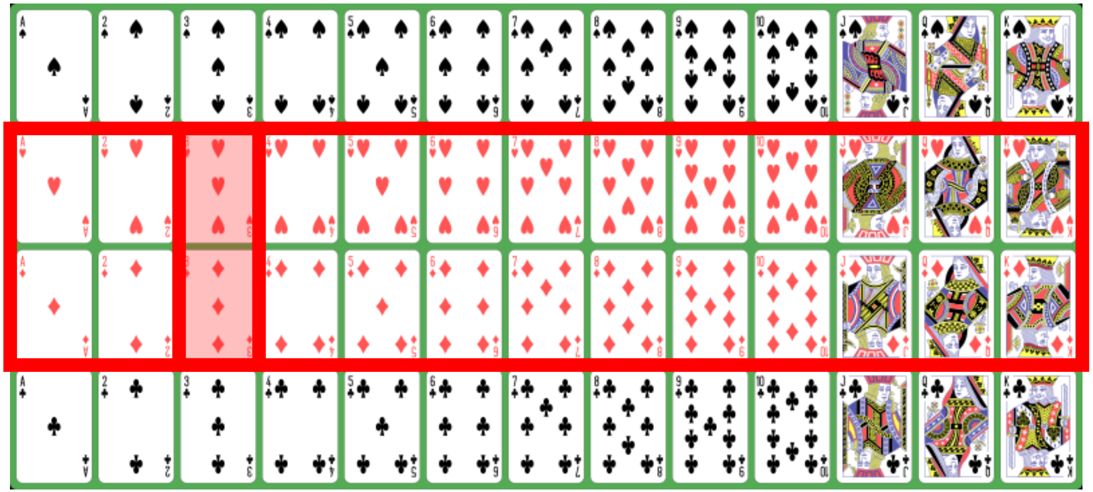
- P(x,y)=P(x∩y)= probability of both x and y occurring



P(red,3) = 2/52 = 1/26

Conditional probability

- $P(x|y) = \frac{P(x \cap y)}{P(y)}$ = probability x occurring, given y



$P(3|red) = (2/52) \ / \ (1/2) = (2*2)/52 = 1/13$

Probability rules

- Sum rule $P(x) = \sum_y P(x \cap y)$
- Product rule $P(x, y) = P(x \cap y) = P(x|y)P(y)$
- Bayes' rule $P(y|x) = \dfrac{P(x|y)P(y)}{P(x)}$
- Chain rule $P(x_1 \cap \cdots \cap x_n) =$

$$P(x_1)P(x_2|x_1)P(x_3|x_2 \cap x_1)\mathrm{P}\left(x_n \middle| \cap_{i=1}^{n-1} x_i\right)$$

Probability terminology

- Prior probability: $P(x)$
  - The probability of x occurring, in general, given no additional information
- Posterior probability: $P(x|y)$
  - The probability of x occurring given that y occurred

Probability terminology

- Independence: no statistical relation between x and y; neither event influences probability of the other
  - $P(x|y) = P(x)$
  - $P(y|x) = P(y)$
  - $P(x,y) = P(x)P(y)$

① no correlated variable
② no one cause the other

- Conditional independence: x and y are independent conditioned on a third variable, z
  - $P(x,y|z) = P(x|z)P(y|z)$

Example: Probability and ML

	Age <18	Age 18-45	Age >45
Purchase = Yes	10	100	50
Purchase = No	90	900	100

P(age>45|yes)
50/160 = **31%**

P(age>45|no)
100/1090 = **9%**

P(yes|age<18)
10/100 = **10%**

P(yes|age18-45)
100/1000 = **10%**

P(yes|age>45)
50/150 = **33%**

- Does knowing customers' ages help you predict purchasing behaviour?
- Does knowing purchase behaviour help you predict customer age?

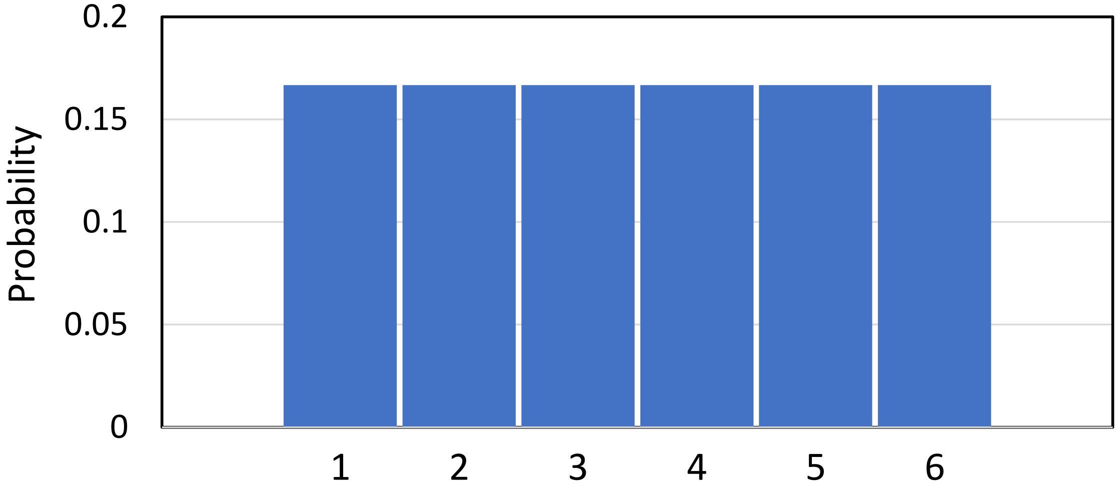
Probability distributions

Probability distribution

- A list of all possible outcomes of a random variable along with their probability values, or a mathematical function that describes the possibilities of different outcomes
- An **empirical** probability distribution is created by observing the frequency of events in the world
- A theoretical probability distribution is based on a mathematical model of a random process

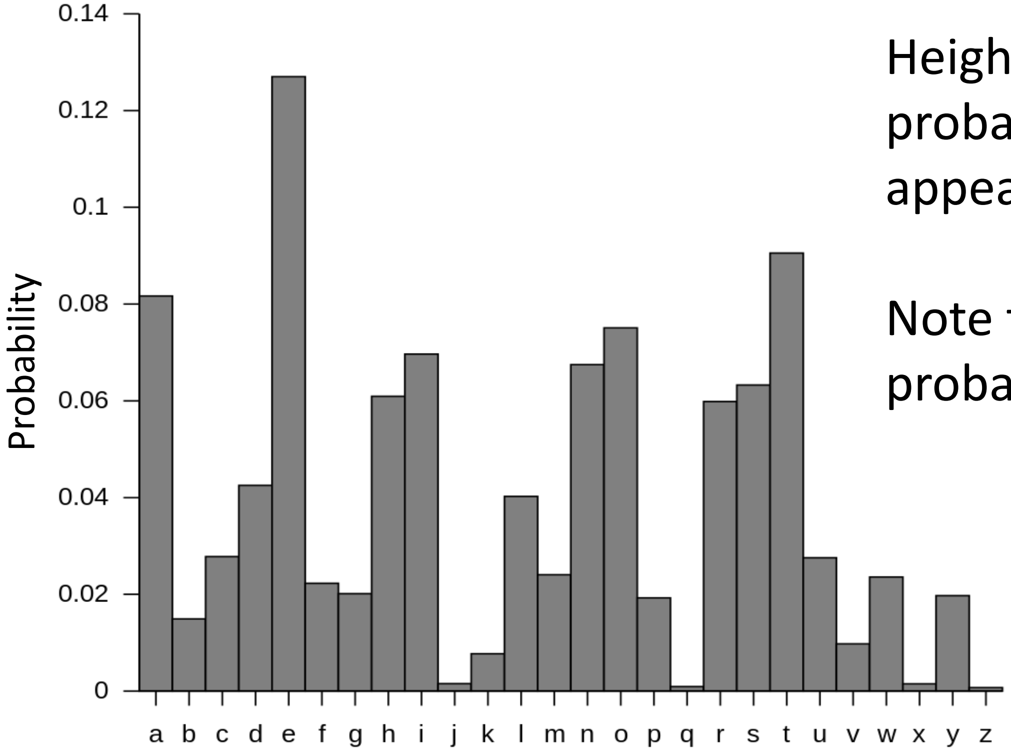
Example: Rolling a die

Outcome	1	2	3	4	5	6
Probability	1/6	1/6	1/6	1/6	1/6	1/6



Discrete uniform distribution = all outcomes equally likely

Example: Letters in English text



Height of bar is the probability of a letter appearing

Note that probabilities sum to 1

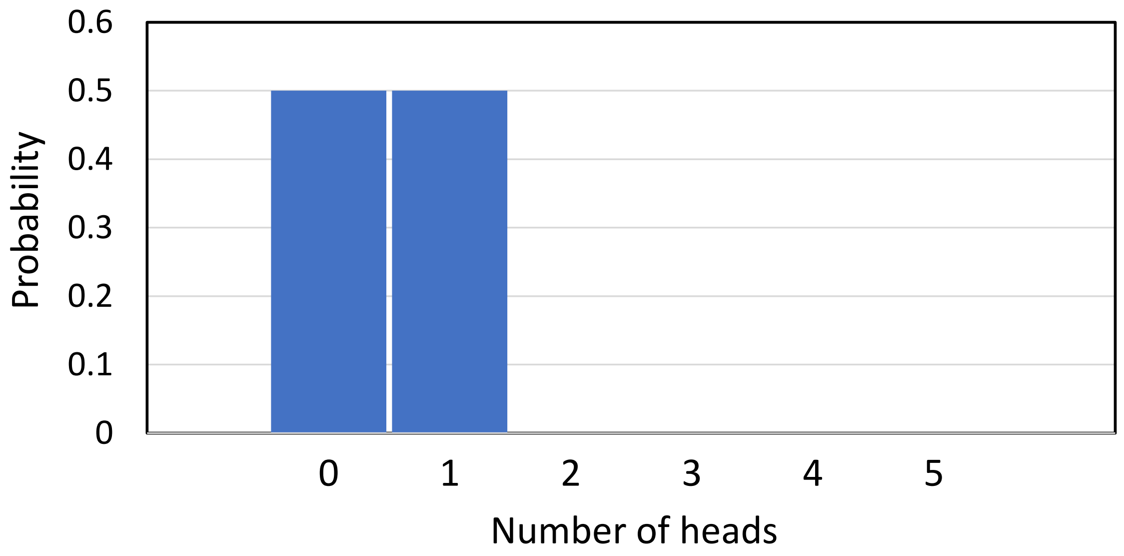
Binomial distribution

- **Bernoulli trials** = independent events with only two possible outcomes
  - Coin flips (heads or tails)
- A **binomial distribution** results from a series of Bernoulli trials
- The probability of an event with probability p occurring exactly m out of n times:

$$B(m;n,p) = \binom{n}{m} p^m (1-p)^{n-m} = \frac{n!}{m!\,(n-m)!} p^m (1-p)^{n-m}$$

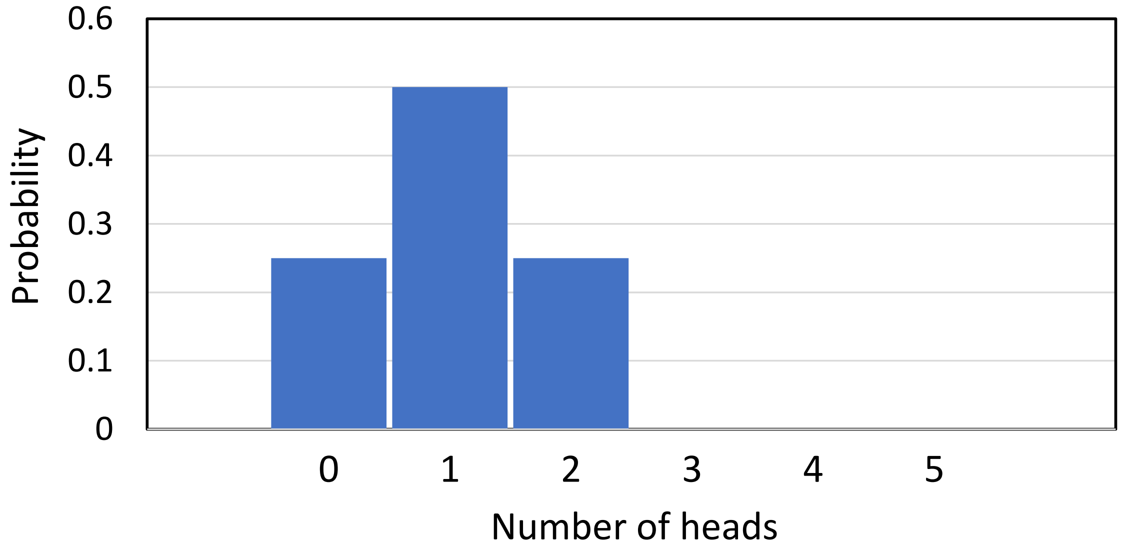
Binomial distribution

- If we flip a (fair) coin once, what is the probability of heads?



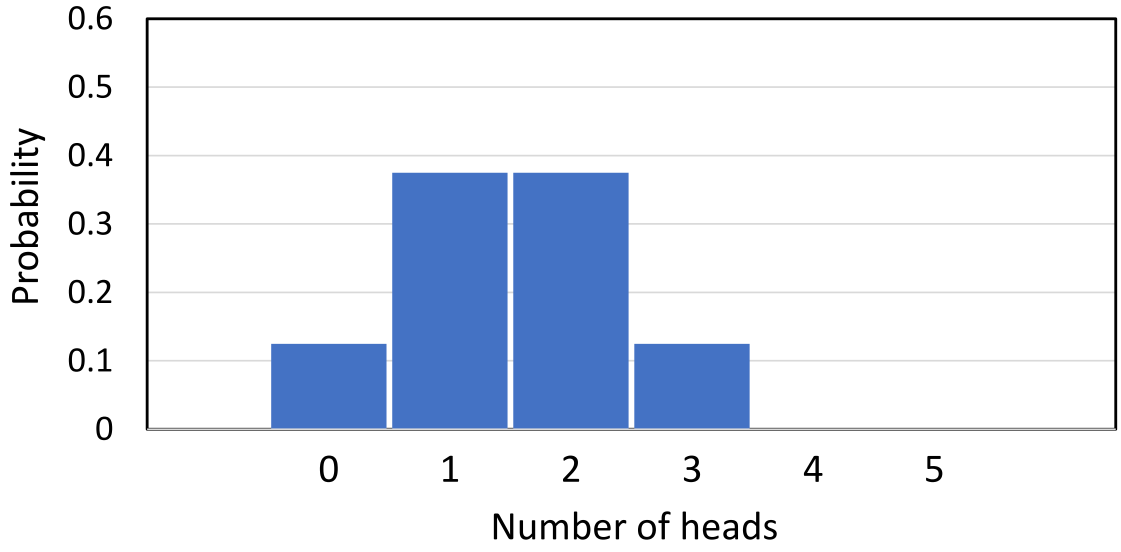
Binomial distribution

- If we flip a (fair) coin twice, what is the probability of 2 heads?



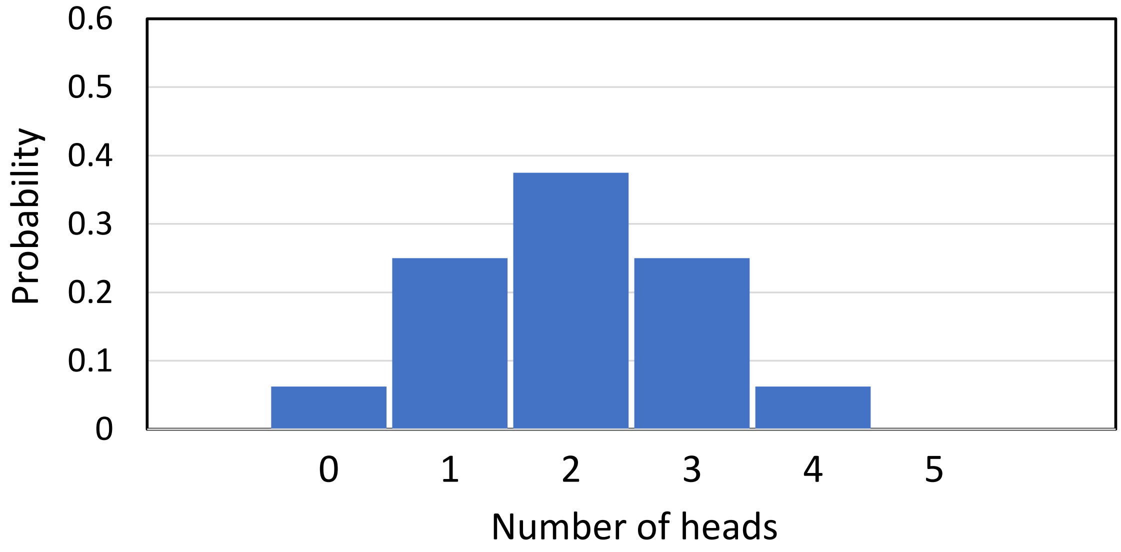
Binomial distribution

- If we flip a (fair) coin three times, what is the probability of 3 heads?



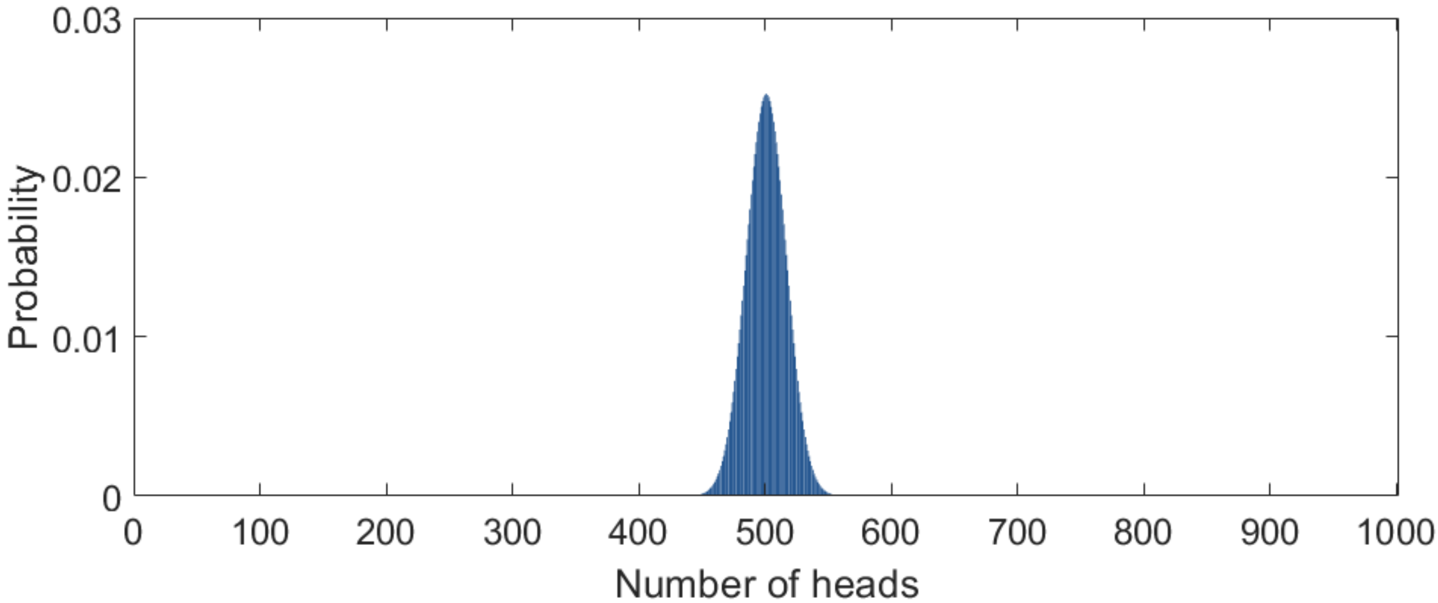
Binomial distribution

- If we flip a (fair) coin four times, what is the probability of 4 heads?



Binomial distribution

- Probability distribution for 1000 coin flips



Multinomial distribution

- A **multinomial distribution** results from a series of independent trials with more than two outcomes
  - Dice rolls (1, 2, 3, 4, 5, or 6)
  - Game outcomes (win, lose, or draw)
- The probability of events X_1, X_2, … X_n with probabilities p_1, p_2, … p_n occurring exactly x_1, x_2, … x_n times, respectively:

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n) = \left(\sum_{i=1}^{n} x_i\right)! \prod_{i=1}^{n} \frac{p_i^{x_i}}{x_i!}$$

multinomial distribution

$$\sum_{i=1}^{n} x_i = n$$

$$\frac{n!}{x_1!\, x_2! \cdots x_n!} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n}$$

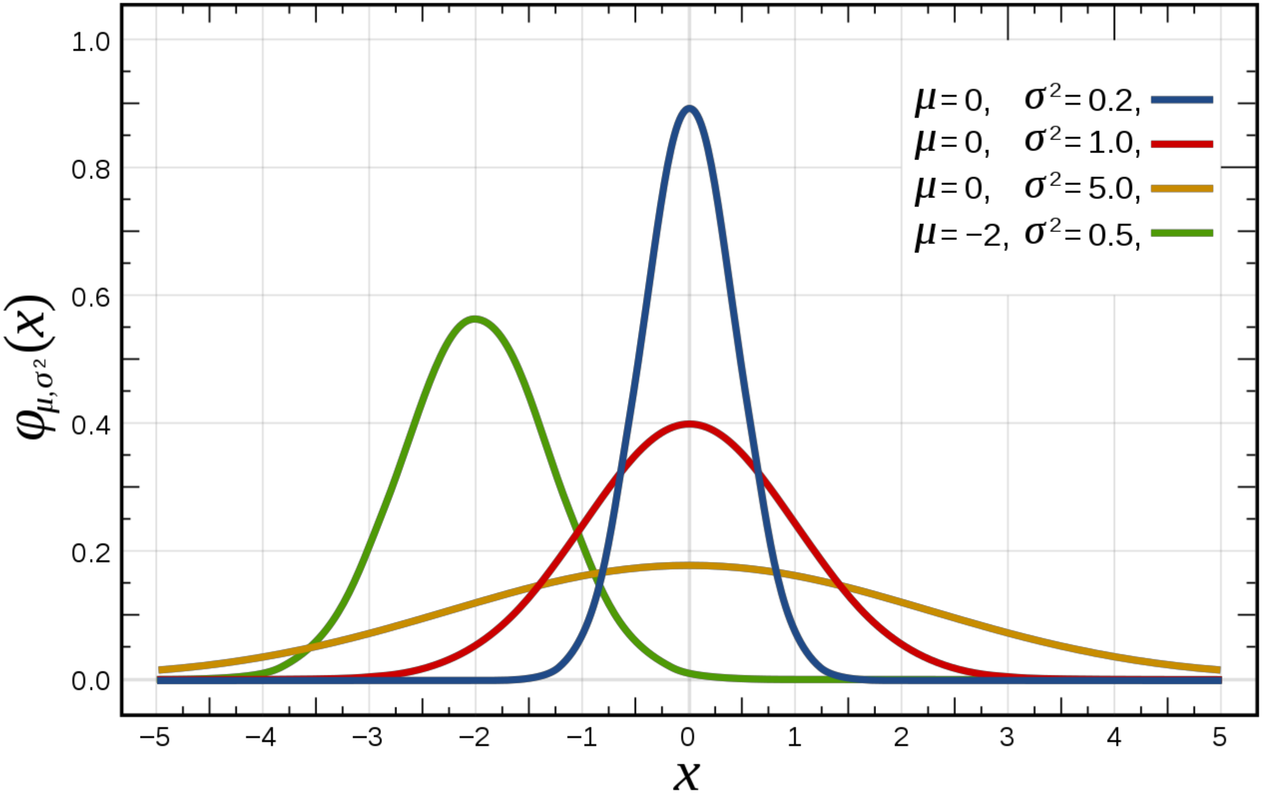
$$= \left(\sum_{i=1}^{n} x_i\right)! \prod_{i=1}^{n} \frac{p_i^{x_i}}{x_i!}$$

Gaussian (normal) distribution

- A **normal distribution** (or **Gaussian distribution**) is often used to represent a noisy continuous variable, when the exact type of noise is unknown
- The probability of observing value x from a variable with mean (expected value) μ and standard deviation σ:

$$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Gaussian (normal) distribution



Probability models

- Probability model = a mathematical representation of a random event
- It consists of
  - A sample space – the set of possible outcomes
  - Events – a subset of the sample space
  - A **probability distribution** of the events
- The model predicts the relative frequency of each event x: P(x)

Examples

- Probability model for a fair coin:
  - P(heads) = 0.5, P(tails) = 0.5

- Probability model for a fair die:
  - P(1) = ? , P(2) = ? , … P(6) = ?

- Probability model for guessing a multiple-choice question with 4 answers:
  - P(correct) = ? , P(incorrect) = ?

Entropy

Why entropy?

- Measure of information
  - How much information is available for learning?
  - How much did the model learn?
  - Are two models representing the same information?

Entropy (information theory)

- (Shannon) **Entropy** is a measure of unpredictability, the information required to predict an event
- Entropy is measured in **bits** (binary digits)
- Entropy of a discrete random variable X with possible states x_1, x_2, … x_n is

$$H(X) = -\sum_{i=1}^{n} P(x_i)\, log_2\, P(x_i) \qquad\qquad 0\, log_2 0 \stackrel{def}{=} 0$$

Entropy

- Entropy of a fair coin flip (P(heads)=0.5)?

$$H(X) = -(P(h)\, log_2\, P(h)) - (P(t)\, log_2\, P(t))$$
$$H(X) = -(0.5\, log_2\, 0.5) - (0.5\, log_2\, 0.5)$$
$$H(X) = -(0.5 * -1) - (0.5 * -1) = 1$$

- Entropy of a trick coin flip (P(heads)=0.9)?

$$H(X) = -(P(h)\, log_2\, P(h)) - (P(t)\, log_2\, P(t))$$
$$H(X) = -(0.9\, log_2\, 0.9) - (0.1\, log_2\, 0.1)$$
$$H(X) = -(0.9 * -0.14) - (0.1 * -0.33) = 0.47$$

Entropy values

- Entropy depends on both the number of possible states, and the likelihood of those states
  - Low entropy = outcomes highly predictable
  - High entropy = outcomes unpredictable
- Range of entropy depends on the number of possible outcomes
  - 2 outcomes: entropy range is 0 – 1
  - N outcomes: entropy range is 0 – log(n)

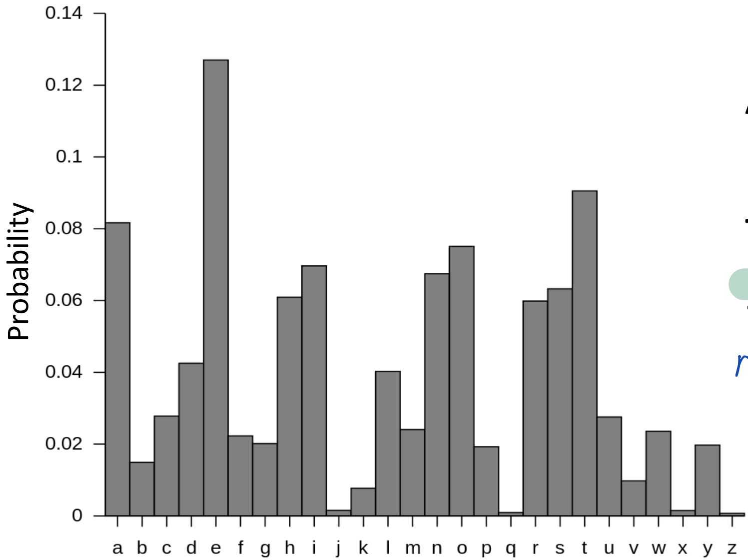
    eg. 4 outcomes
    range 0 - 2
  - Entropy = 0 means only one outcome is possible
  - Entropy = log(n) means all outcomes equally likely

$$H(X) = -\sum_{i=1}^{n} \frac{1}{n} \log\left(\frac{1}{n}\right) = -n \cdot \frac{1}{n} \cdot \log\left(\frac{1}{n}\right) = \log(n)$$

Entropy and message encoding

- What's the entropy of English text?
  - 26 letters = 26 outcomes = log(26) = 4.70 bits

    total = 4.7 x no. of letter
  - …assuming every letter is equally likely to appear



Actual entropy = 4.14

Minimum letters needed to achieve this entropy:

$2^{4.14} = 17.63$

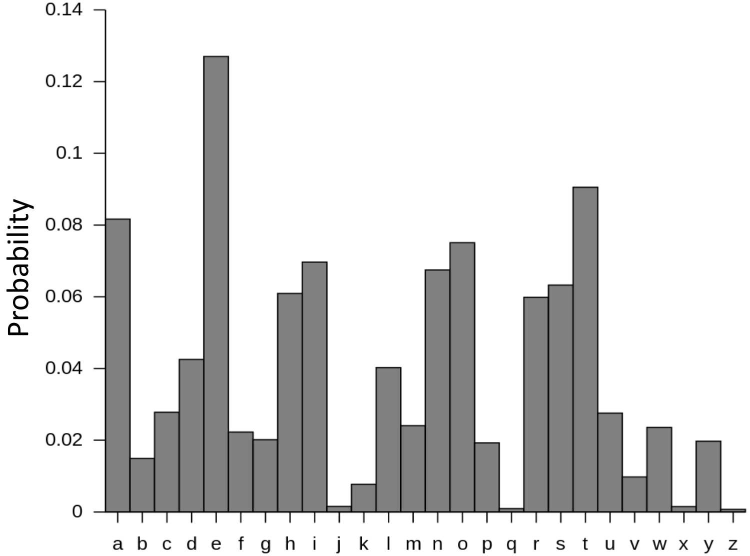
n outcomes $0 - \log_2(n)$.

$\log_2(n) = 4.14 \rightarrow n = 2^{4.14}$

Shannon, C. E. (1951). Prediction and Entropy of Printed English. *The Bell System Technical Journal*, 30(1), 50-64.

Entropy and compression

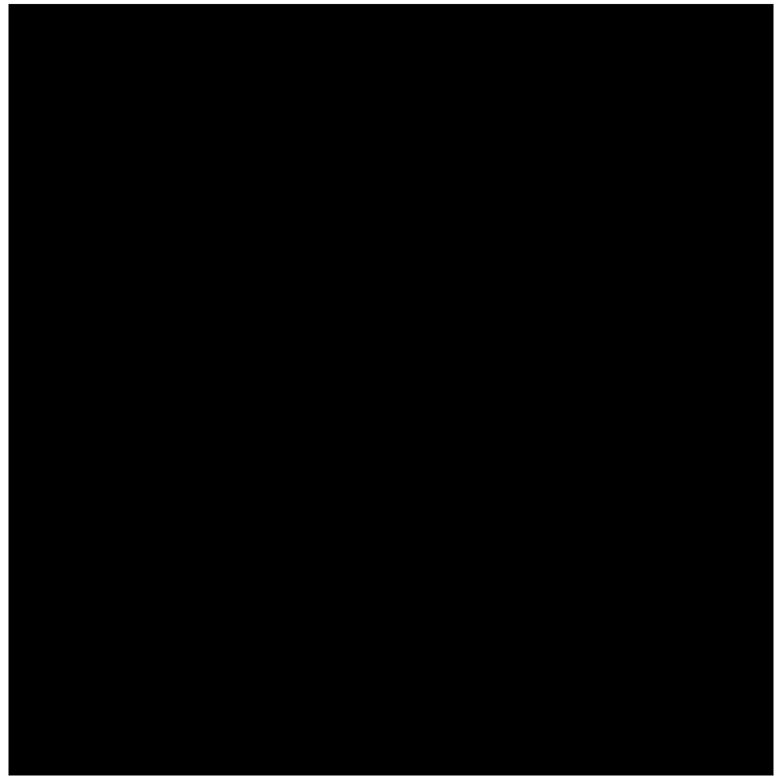
- To send a message with the minimum possible bits, assign shorter codes to the most frequent letters



Entropy and encoding

- How to pack maximum information into limited bandwidth?
- Example: a camera with 1 bit per pixel

Threshold so 0% of pixels are white
P(0) = 1, P(1) = 0
Entropy = 0



Entropy and encoding

- How to pack maximum information into limited bandwidth?
- Example: a camera with 1 bit per pixel

Threshold so 1% of pixels are white
P(0) = 0.99, P(1) = 0.01
Entropy = 0.08

Entropy and encoding

- How to pack maximum information into limited bandwidth?
- Example: a camera with 1 bit per pixel

Threshold so 10% of pixels are white
P(0) = 0.9, P(1) = 0.1
Entropy = 0.47

Entropy and encoding

- How to pack maximum information into limited bandwidth?
- Example: a camera with 1 bit per pixel

Threshold so 50% of pixels are white
P(0) = 0.5, P(1) = 0.5
Entropy = 1
most informative

Entropy and encoding

- How to pack maximum information into limited bandwidth?
- Example: a camera with 1 bit per pixel

Signal with higher entropy conveys more information

more entropy → more information

Example: Entropy and ML

Temperature (T)	Play?
6	no
7	no
10	no
14	yes
17	yes
18	yes
19	yes
22	no
23	yes
24	yes

- What threshold gives the most information about whether or not to play outside?

T=12

Entropy = 0 | 0.59

low entropy

T=16

Entropy = 0.97 | 0.72

T=20

Entropy = 0.99 | 0.92