BSCS 2019 - Neural Computation

III - Probabilistic models

Mihály Bányai
banyai.mihaly@wigner.mta.hu
http://golab.wigner.mta.hu/people/mihaly-banyai/

- Random variables and distributions
- Building and handling graphical models
- Continuous random variables

- Random variables and distributions
- Building and handling graphical models
- Continuous random variables

Probability as information

- We are using probability theory to quantify the uncertainty of observed quantities
- We also want to make inferences about quantities we do not observe directly
    - e.g. from a visual observation (retinal activation) the brain wants to infer what kind of objects are there in the environment
    - for this, we need a model that tells us how different object alignments produce different retinal activations
    - if we can formalise this forward mapping in a model, we can also do the inverse calculation in it

"Whether you can observe a thing or not depends on the theory which you use. It is the theory which decides what can be observed."

Albert Einstein

Recommended reading

http://www.johndcook.com/blog/2014/01/21/probability-is-subtle/

"To understand God's thoughts we must study statistics, for these are the measure of His purpose."

Florence Nightingale

If I flip a coin, look at the result, but don’t show you, its state is random for you but not for me. Probability describes knowledge.

Random variables

- A random variable *x* is a quantity of unknown value, e.g. a grade of a student (this is not the formal definition)
- each possible value of the random variable corresponds to a proposition
  - X_1 = “The grade is 1”, X_2 = “The grade is 2”, etc.
    - one of these proposition is also called a sample of the random variable
- based on our axiom set, we derive the probabilities of these propositions
- the probabilities of all possible values of *x* define its **probability distribution**

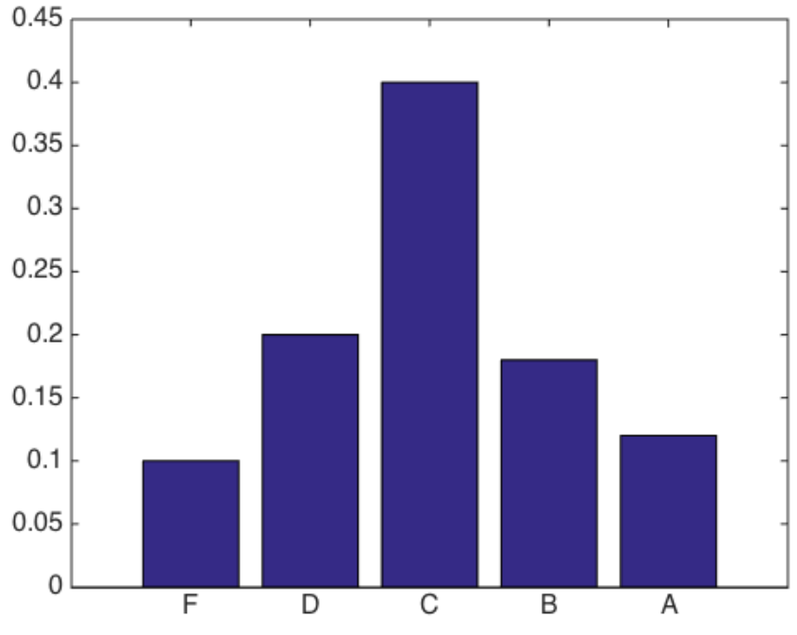
Probability distributions

- the axiom set contains two types of propositions: knowledge about the world and assumptions

  - AS = {“The probability of a student getting a grade 1 in a science class is 0.2 ”,
    “The probability of a student getting a grade 2 in a science class is 0.3 ”,
    …
    “The probability of a student getting a grade 1 in a PE class is 0.1 ”,
    …
    “The probability of a class being a science class is 0.3”,
    “The probability of a class being a PE class is 0.7”,
    …}

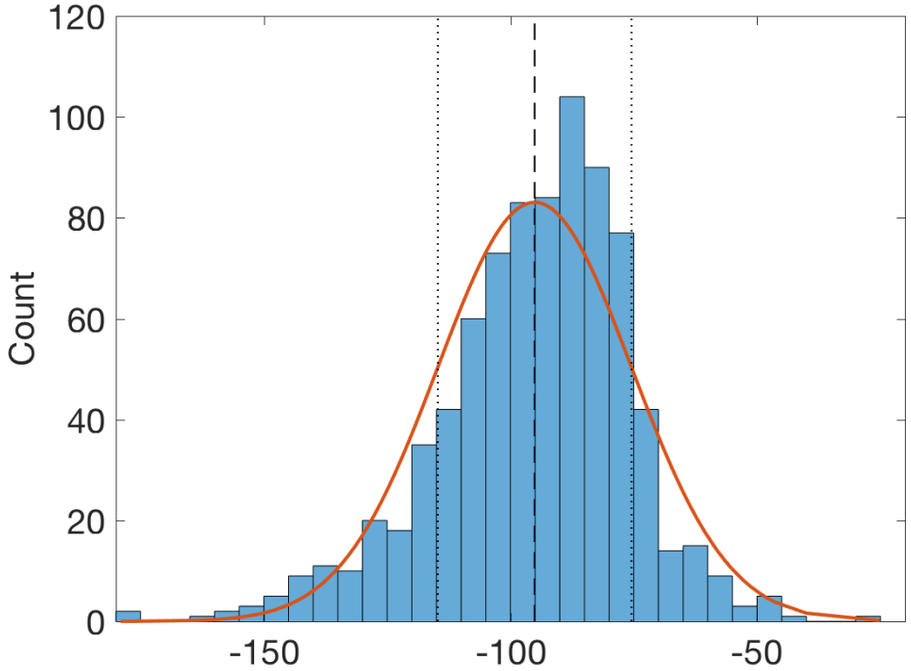
- by the product and sum rules:
  $Pr(X_1) = 0.3 \times 0.2 + 0.7 \times 0.1 = 0.13$

- if we do this for all possible values of *x,* we obtain the **probability mass function (PMF),** P(*x)*



Mean and variance

- If we want to give a concise description about a random variable (possibly with very many values), we can give an **average** value of it, and a measure of how much are the actual values typically **dispersed** around this average



- the mean (expectation) of the variable is given as a sum of all possible values weighted by their probabilities
  - $E(x) = \sum_x x \, P(x)$
  - the mean grade is 0.2x1 + 0.3x2 + …
- the variance of the variable is given as the expectation of the squared deviation from the mean
  - $Var(x) = E((x - E(x))^2)$

Joint probability distributions

- P(x,y) is the **joint PMF** of two variables, which tells us Pr(x=i ∧ y=j) for each i and j in the possible value sets of the two variables

- the two **marginal PMFs**, P(x) and P(y) can be obtained by the sum rule (marginalisation)

joint PMF of the outcome combinations of two dice

$i \backslash j$	1	2	3	4	5	6	$p_X(i)$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p_Y(j)$	1/6	1/6	1/6	1/6	1/6	1/6	

Example: Joint Distribution of S and M.
S = The sum of two dice, M = The maximum of two dice.

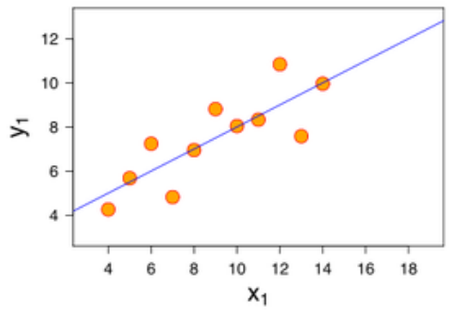
	b						$p_S(a)$
a	1	2	3	4	5	6	
2	1/36	0	0	0	0	0	1/36
3	0	2/36	0	0	0	0	2/36
4	0	1/36	2/36	0	0	0	3/36
5	0	0	2/36	2/36	0	0	4/36
6	0	0	1/36	2/36	2/36	0	5/36
7	0	0	0	2/36	2/36	2/36	6/36
8	0	0	0	1/36	2/36	2/36	5/36
9	0	0	0	0	2/36	2/36	4/36
10	0	0	0	0	1/36	2/36	3/36
11	0	0	0	0	0	2/36	2/36
12	0	0	0	0	0	1/36	1/36
$p_M(b)$	1/36	3/36	5/36	7/36	9/36	11/36	1

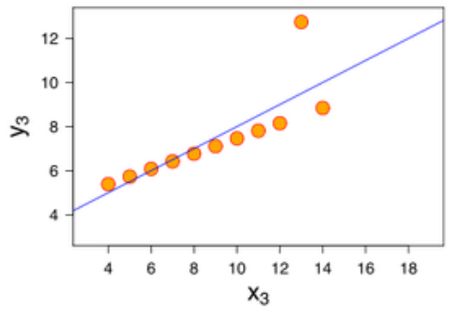
Correlation

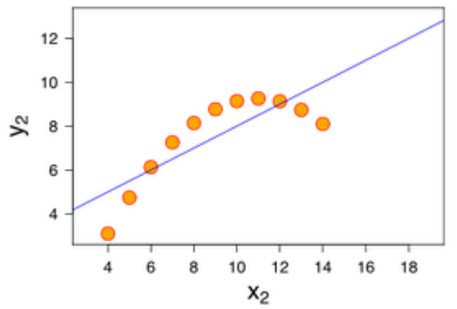
- If we have two (or more) random variables, additional to their mean and variance, we can characterise them by measuring how much they typically move in the same direction, i.e. when one is bigger the other is also bigger

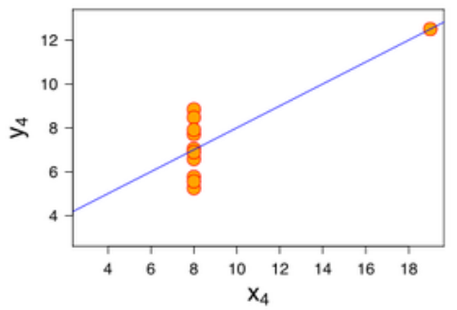
  - Correlation describes how well can a linear function describe the relationship between the two variables

  - Correlation, similarly to anything else in probability theory, does not tell us anything about causal relationships between the variables





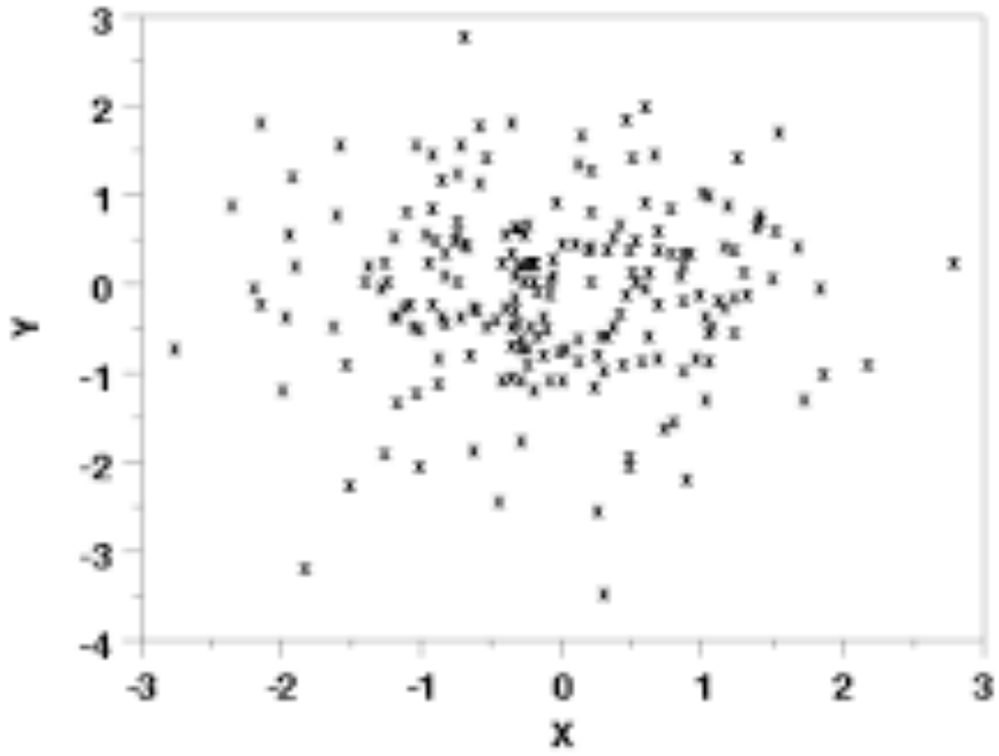




Recommended reading
http://www.tylervigen.com/spurious-correlations

Independence

- independence of two variables means that their joint PMF equals to the product of the two marginal PMFs
  - $x \perp y \Leftrightarrow P(x,y) = P(x)P(y)$
  - if I have two dice, the probability of getting two sixes is 1/36, exactly the product of getting one six with one of the cubes and another six with the other



- this also means that the value of one variable does not contain any information regarding the value of the other, so conditioning on it does not change the PMF
- $x \perp y \Leftrightarrow P(x|y) = P(x) \Leftrightarrow P(y|x) = P(y)$

Conditional independence

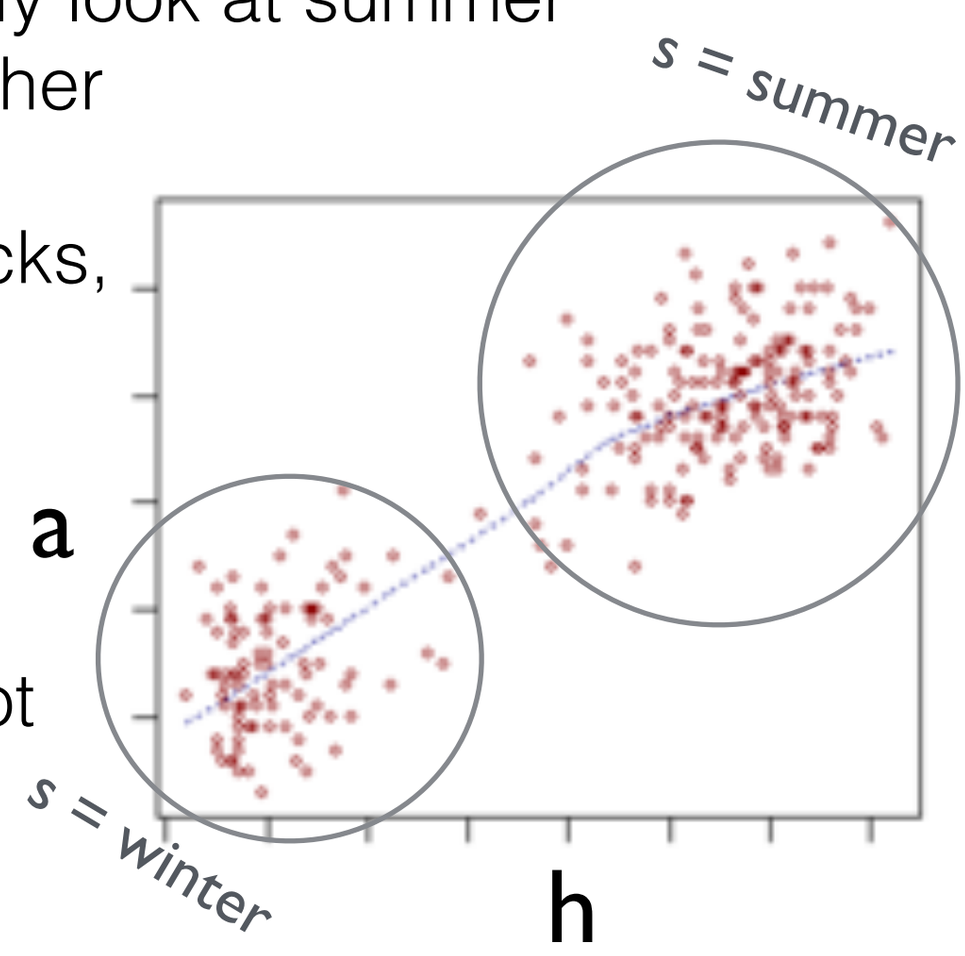
- conditional independence means that two variables are only independent if we know the value of a third

  - occurrence rates of hiking trips and shark attacks on a given day are clearly not independent - but that's because both are more likely to happen in the summer. If I only look at summer periods, within those they do not vary together

  - h = no. of hiking trips, a = no. of shark attacks, s = season

  - $h \perp a|s \Leftrightarrow P(h,a|s) = P(h|s)P(a|s)$

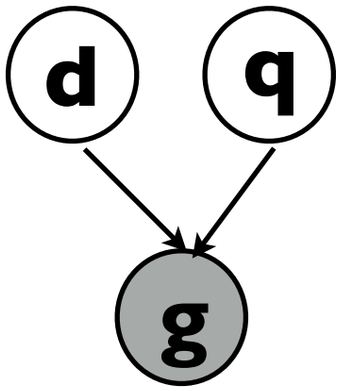
- marginal and conditional independence do not follow
  from each other



- Random variables and distributions
- Building and handling graphical models
- Continuous random variables

Building a probabilistic model

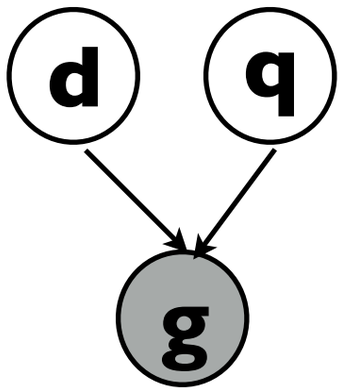
- We will also use a **graphical** representation of the model that will help our intuition
- first, we collect the variables that we want to talk about in our model
- what will be the quantity that we will be able to observe?
  - grade of a test, g ∈ {A,B,C,D,F} = [5,4,3,2,1]
  - we add a shaded circle to the graph with the **observed variable**



- what quantities do we want to infer?
  - preparedness of a student, q ∈ {low,medium,high} = [1,2,3]
  - we add an empty circe with the **unobserved variable** (also called **latent** or **hidden**), and connect it to the observed one with an arrow - this means that they are dependent
- might there be any other variables that influence how these two quantities relate to each other?
  - difficulty of the test, d ∈ {easy,moderate,hard} = [1,2,3]
  - we add this hidden variable to the graph as well in a way that it is directly connected to the grade, but not preparedness (the difficulty of a test and the preparedness of a person who did not create it are independent from each other)

Meaning of the graphical model

- The graphical model is the skeleton of our probabilistic model - it does not define any probabilities yet, only **independence** relations between variables.
  - without any such information, we have to assume that all variables depend on all others
  - in this example, d and q are independent:
    - $d \perp q \Leftrightarrow P(d,q) = P(d)P(q) \Leftrightarrow P(d|q) = P(d) \Leftrightarrow P(q|d) = P(q)$

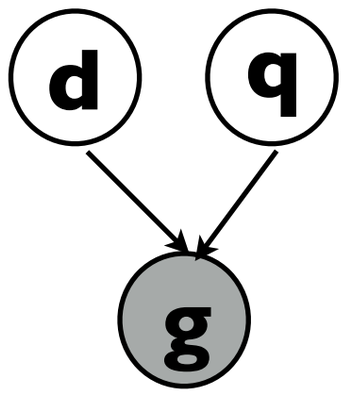


- The graphical model implies a **factorisation** of the joint probability mass function of all our variables in which every variable is conditioned on its parents.
- $P(g,d,q) = P(g|d,q)\, P(d)\, P(q)$
  - in fact, the factorisation can be obtained by the (repeated) application of the product rule, taking conditional independence into account
  - $P(g,d,q) = P(g|d,q)\, P(d,q) = P(g|d,q)\, P(d|q)p(q) = P(g|d,q)\, P(d)\, P(q)$

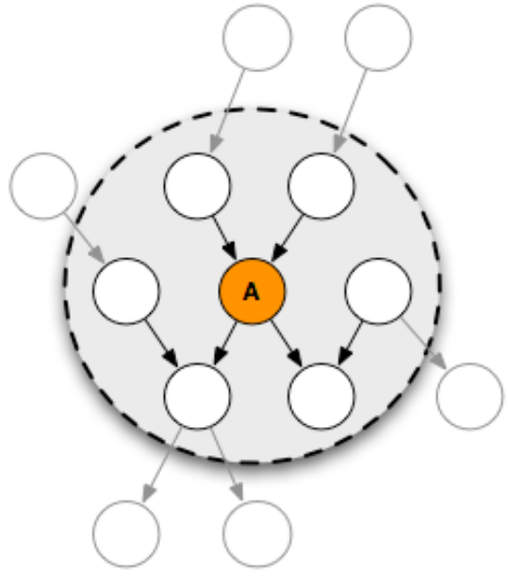
> "There is a santa claus called joint probability distribution."
>
> *Judea Pearl*

Independence relations in graphical models

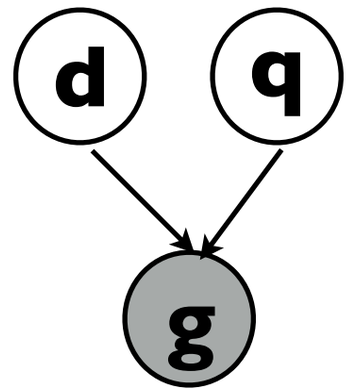
- The observation of a variable can change independence relations
  - since marginal and conditional independence are not the same
  - preparedness and test difficulty are independent *a priori* - but given a specific grade, the are not - a phenomenon called **explaining away** (a v-shape in graphical models)
    - if the grade is D, what would you think about the probabilities of different preparedness values if the test was hard, and what if it was easy? - hardness of the test *explains away* the bad grade, you don't need the preparedness to be low to make it probable



- Any variable in a graphical model is independent of the rest conditioned on:
  - its parents
  - its children
  - the children's other parents



Finishing up the probabilistic model



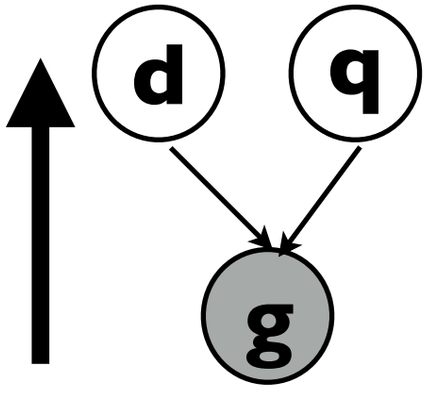
- We have to define the conditional probabilities we factored the joint distribution to.
  - P(g,d,q) = P(g|d,q) P(d) P(q)
- We need 3 PMFs
- We derive the PMF of the hidden variables using the axiom set, encoding our knowledge about the world and assumptions about the phenomena at hand
  - e.g. we know that medium preparedness is twice as likely as low or high: P(q) = [0.25 0.5 0.25]
  - e.g. we assume that all test difficulties are equally probable: P(d) = [0.33 0.33 0.33]
- We have to quantify the PMF of g using the axioms and the value combinations of the two additional conditions
- all possible combinations of the conditions will imply a different PMF for g
- We can do this with conditional probability tables

low prep. easy test | low prep. med. test

Grade	1,1	1,2	…
1	0.1	0.2	
2	0.3	0.4	
3	0.4	0.3	
4	0.1	0.08	
5	0.1	0.02	

Inference in probabilistic models

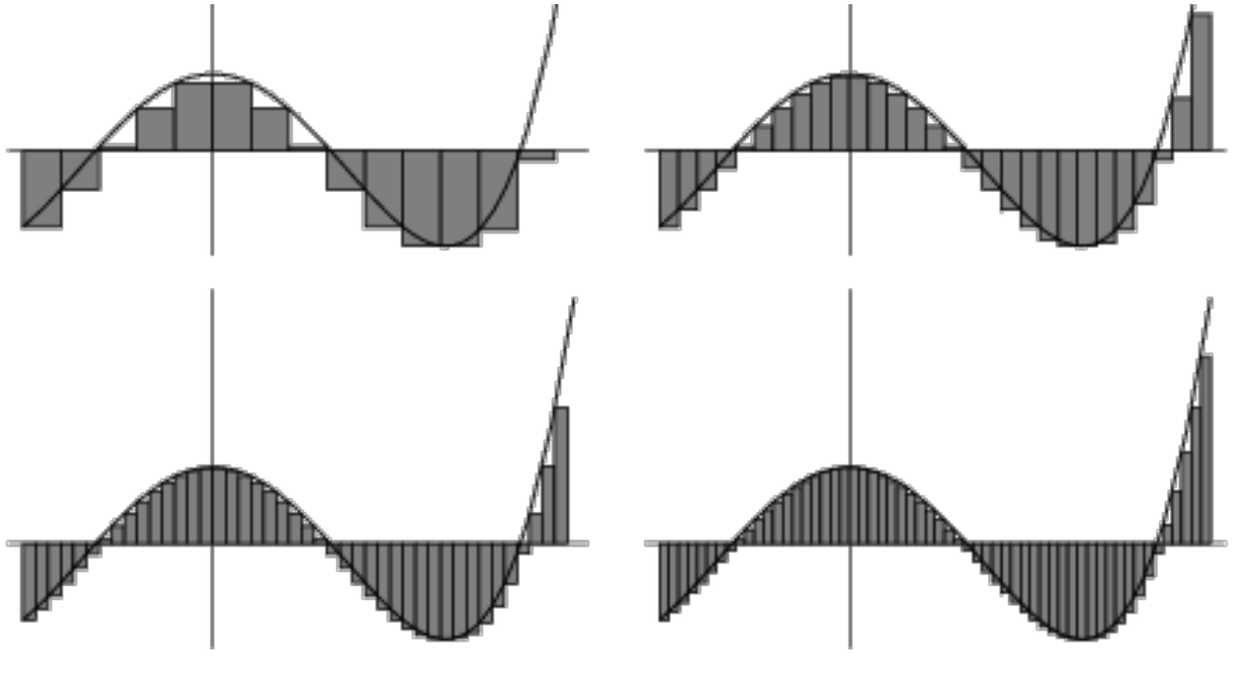
- we want to know how probable different values of the hidden variables are given a value of the observation
- by repeated application of the sum and product rules, the Bayes theorem and conditional independence identities, we can deduce these values
- let's say we want to know how probable is it that the student's preparedness is high if we see an 'A' grade: P(q=3|g=5)
  - sum rule: $P(q|g) = \sum_d P(q,d|g)$
  - Bayes theorem: $P(q,d|g) = P(g|q,d)\,P(q,d) / P(g)$
  - sum rule: $P(g) = \sum_d \sum_q P(g,d,q)$
    - product rule $P(g) = \sum_d \sum_q P(g,d,q) = \sum_d \sum_q P(g|q,d)\,P(q,d)$
    - independence $P(g) = \sum_d \sum_q P(g,d,q) = \sum_d \sum_q P(g|q,d)\,P(q)\,P(d)$
  - substituting in: $P(q|g) = \sum_d P(g|q,d)\,P(q)\,P(d) / \sum_d \sum_q P(g|q,d)\,P(q)\,P(d)$
  - we arrived to an expression for the conditional PMF we were looking for that only contains PMFs we defined in the model, so we can evaluate this without problem



- Random variables and distributions
- Building and handling graphical models
- Continuous random variables

Technical note: summing infinitely many values

- sometimes we want to calculate with variables defined over a continuous interval instead of some discrete values - e.g. the height of a person
- for this we need a generalisation of summation
- we divide the interval $x \in [a,b]$ of the function $f(x)$ to smaller intervals, and draw bars over them just touching the value of the variable
- we make the divisions finer and finer, thus approximating the area under the curve
- this is called an **integral** with the following notation:

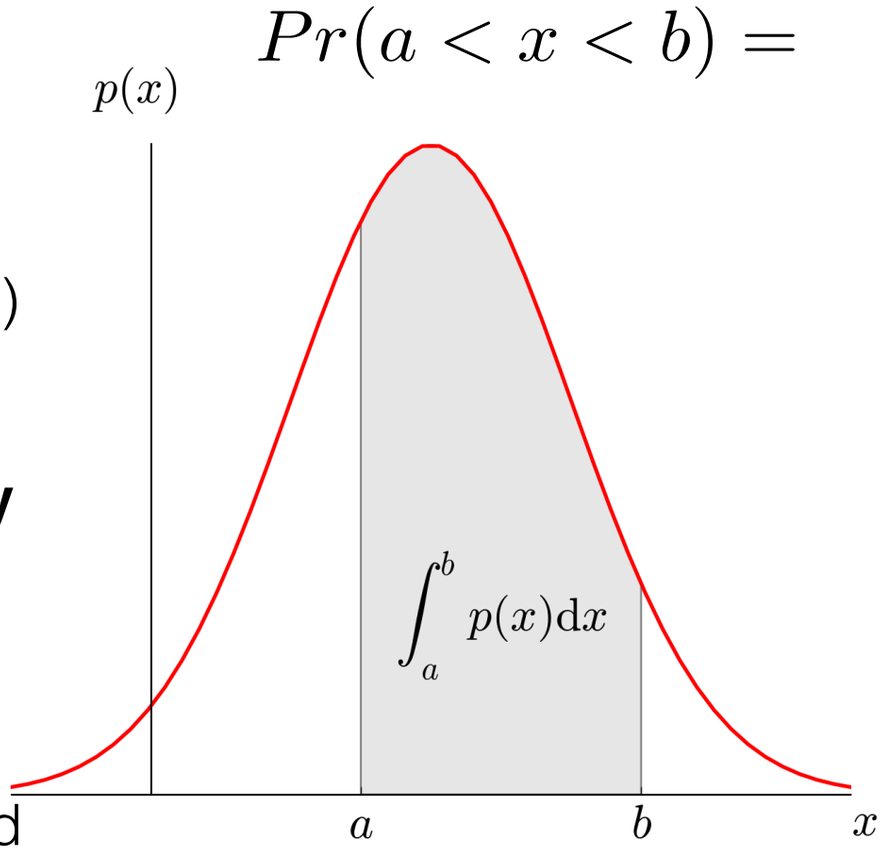


$$\int_a^b f(x)\mathrm{d}x$$

Recommended reading
http://platonicrealms.com/encyclopedia/zenos-paradox-of-the-tortoise-and-achilles

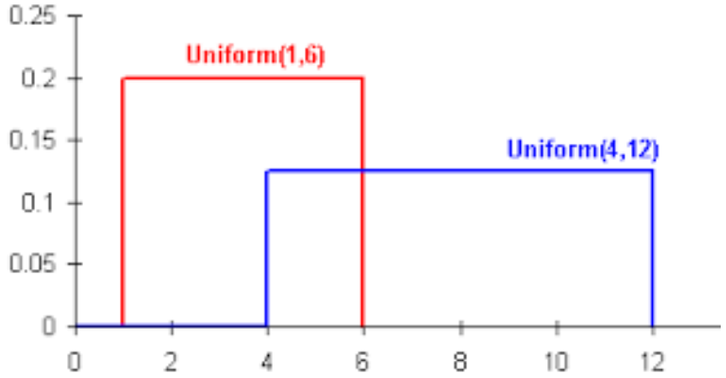
Continuous random variables

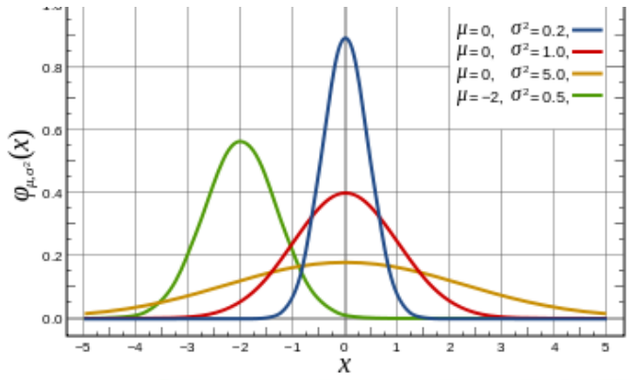
- infinitely many values
- the problem is that each exact value has zero probability (e.g. what is the probability that someone is exactly 1.825375317… meters tall?)
- intervals on the other hand have nonzero probabilities (e.g. the probability that someone is between 1.7 and 1.8 meters)
- we can define a function that gives us the probability of the intervals, conditioned on the axioms - called the **probability density function (PDF),** *p(x)*
- we have to know (or assume) something about the variable to choose a density, e.g. an interval that contains all possible values or a typical value and the dispersion around it
- if we don't know or assume anything then the probability is not quantifiable - if we don't have sufficient prior knowledge and we are not comfortable with assumptions, this is the case

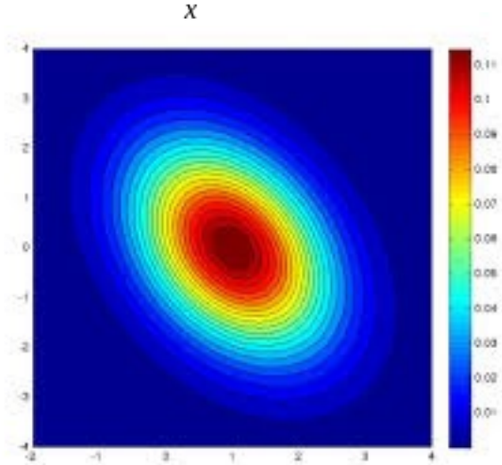


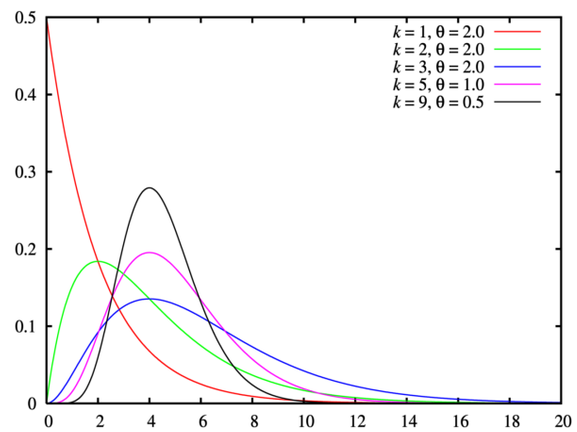
Technical note - parametric density functions

- **Uniform** - U(x; a,b) = $\begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$
  - bounded
  - parameters: a - minimum, b - maximum
- **Gaussian** (normal): symmetric unbounded
  - one dimensional: N(x;μ,σ) = $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
    μ - mean, σ - standard deviation
  - multidimensional:
    N(x;μ,Σ) = $(2\pi)^{-\frac{k}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$,
    μ - mean, Σ - covariance matrix
- **Gamma**: Gam(x;k,θ) = $\frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$
  - positive
  - k - shape, θ - scale









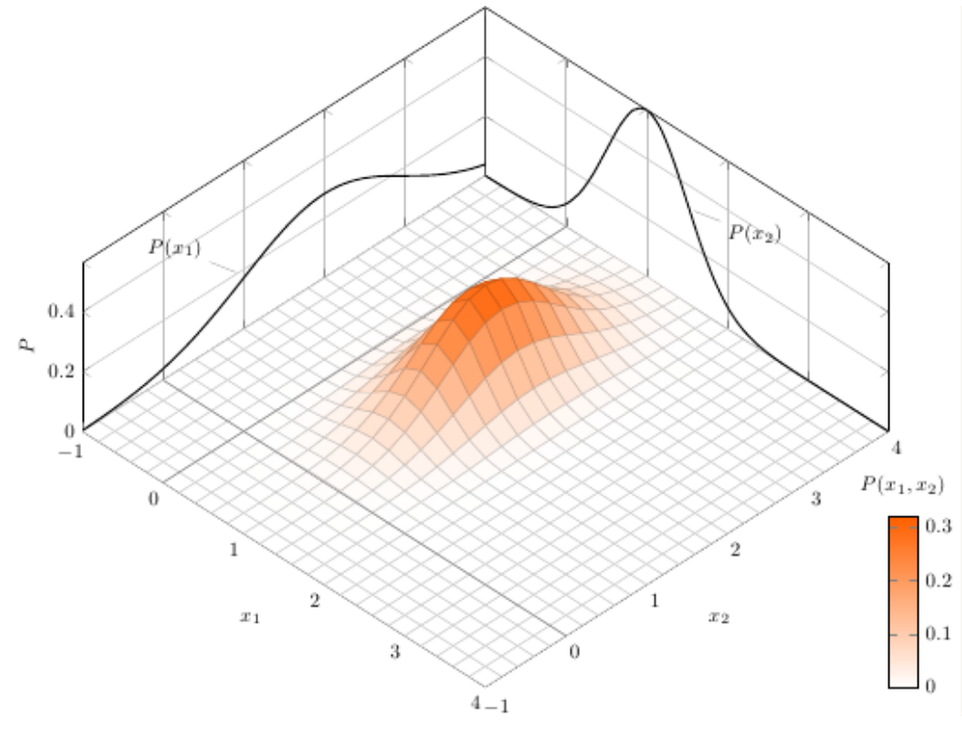
Sum rule on continuous variables

- a two-dimensional PDF is a *surface*

- the summation becomes an integral over all possible values of one of the two variables

- the product rule and criteria for independence remain the same

$$p(x) = \int_{-\infty}^{\infty} p(x, y) \mathrm{d}y$$



$$x \perp y \equiv p(x, y) = p(x)p(y)$$

$$x \perp y \mid z \equiv p(x, y \mid z) = p(x \mid z)p(y \mid z)$$

Bayes theorem for continuous variables

posterior

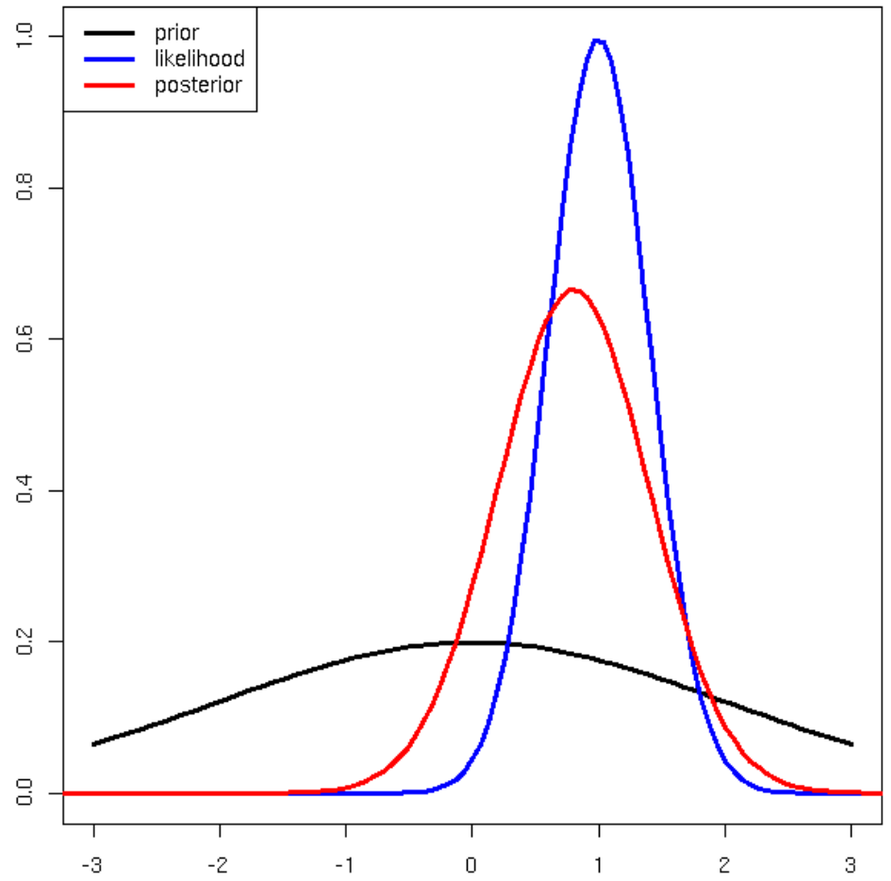
likelihood

prior

$$p(x \mid y, z) = \frac{p(y \mid x, z)p(x \mid z)}{p(y \mid z)}$$

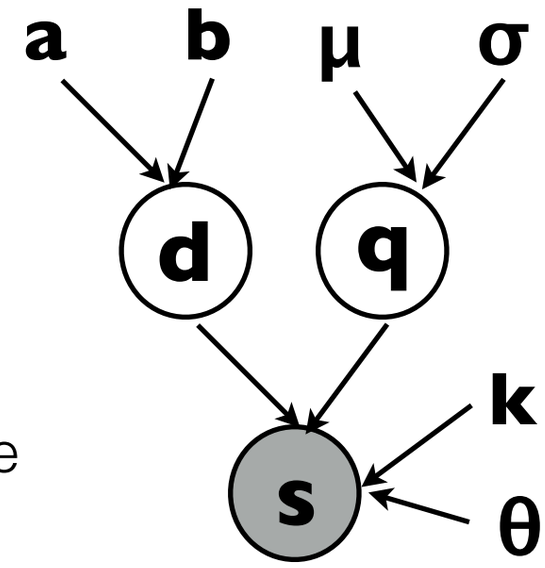
evidence or marginal likelihood

$$p(y \mid z) = \int_{-\infty}^{\infty} p(y \mid x, z)p(x \mid z)\mathrm{d}x$$



Graphical models with continuous variables

- instead of the grade, we want to model the test score
- the intelligence is measured in IQ, p(q) = N(q; 100,15)
- the difficulty is a uniform variable, p(d) = U(d;1,10)
- the conditional PDF of the score can be a Gamma variable, where the scale parameter (the location of the maximum) depends on the hidden variables
  - p(s|d,q) = Gam(s;2,f(d,q))
  - the dependence would be something proportional to the IQ and inversely proportional to difficulty
    e.g. f(d,q) = q (1/d)
- we denote PDF parameters in the graphical model with letters without circles



Inference with continuous variables

- we want to know how probable it is that the student has an IQ over 120 if we observed a test score of 100, that is Pr(q>100|s=100) = $\int_{120}^{\infty} p(q \mid s = 100)\mathrm{d}q$
- analogously to the discrete model, we apply the sum, product, Bayes and independence rules in the same order, obtaining a similar expression

$$p(q \mid s) = \frac{\int_1^{10} p(s \mid q, d)\mathrm{d}d}{\int_1^{10}\int_{-\infty}^{\infty} p(s \mid q, d)p(q)p(d)\mathrm{d}q\mathrm{d}d}$$

- in this formula we again only have PDFs we defined in the model

- but contrary to the discrete case, when substituting in the PDF formulas, we can easily arrive to an expression that we cannot evaluate exactly due to the difficult integrals - in this case we need to use some approximation

> "An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."
>
> *John W. Tukey*

Probabilistic programming languages

- You can define probabilistic generative models in them, supply data and run inference automatically over hidden variables
- Stan - inference by sampling the posterior http://mc-stan.org/
- Edward & Pyro - fuse PPLs with deep learning http://edwardlib.org/
- Church - introduction of a Turing-complete PPL, focus on cognitive modelling https://probmods.org/
- BUGS - the first widely used PPL, now sort of outdated

The way forward

- now we have a toolset to define representations of quantities related to observations
    - we can formalise our knowledge about how they are related to each other
        - we can make inference about non-observed quantities from the observed ones using the model
- we can move on to formulate and test predictions about perceptual problems using probabilistic models of the stimuli, assuming that the brain also tries to use such representations
- then we have to tie the model variables and inference algorithms to neurons