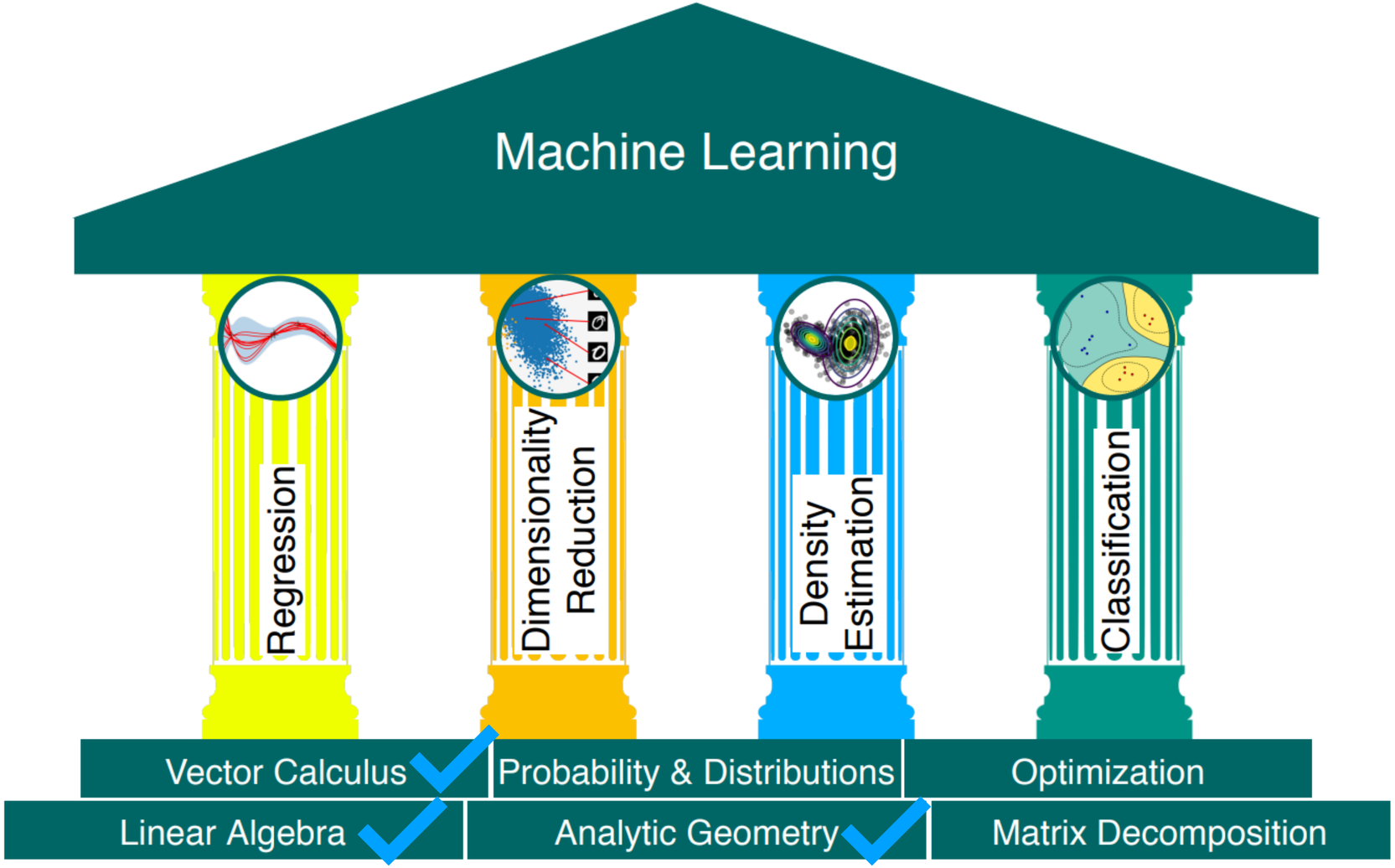
Probability and Distributions

Week 6 - Introduction to ML / Thang Bui / ANU / 2023 S2

Who am I and where is Jo?

Foundations of ML



Source: Maths for ML Textbook

Probability examples and why we need to care

Canberra Forecast

No warnings for the Australian Capital Territory

Forecast updated at 9:33 am EST on Tuesday 15 August 2023.

Source: BOM 15 August 2023

Forecast for the rest of Tuesday

Max 14

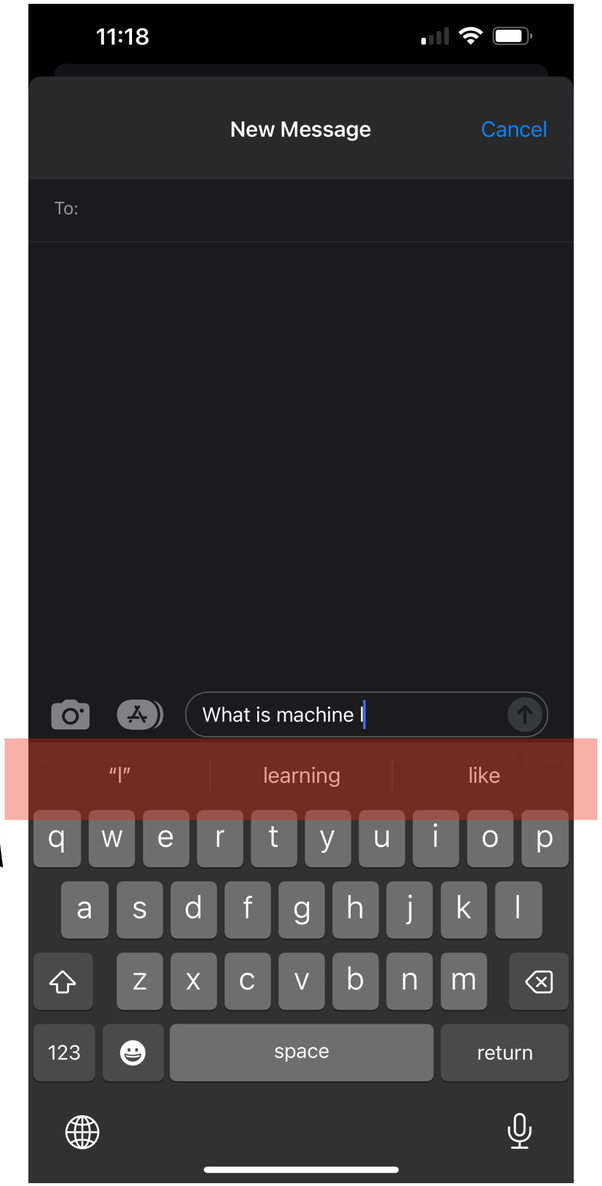
Partly cloudy.

Chance of any rain: **20%**

Canberra area

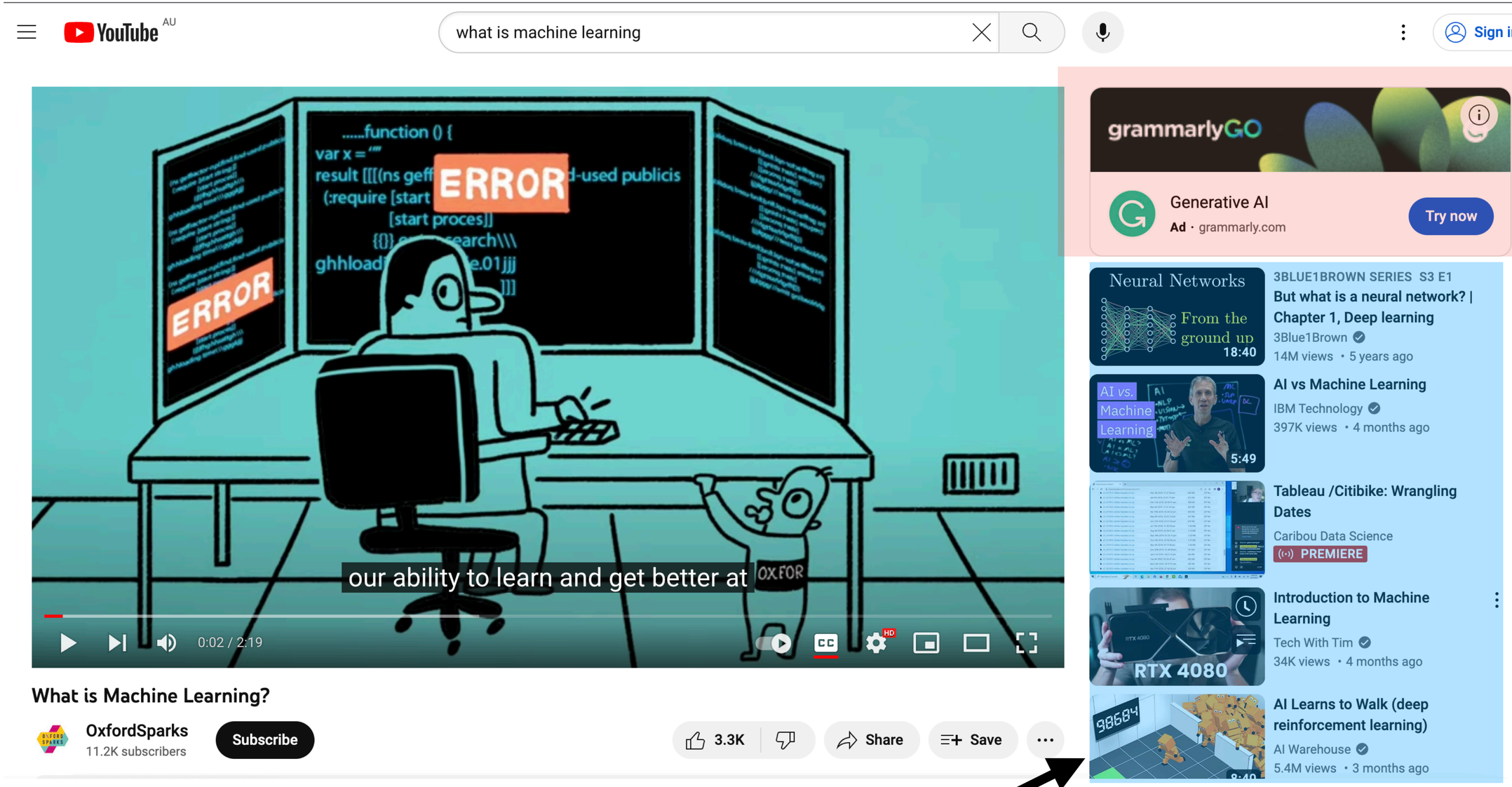
Partly cloudy. Light winds.

What is the probability that it will rain today or tomorrow?



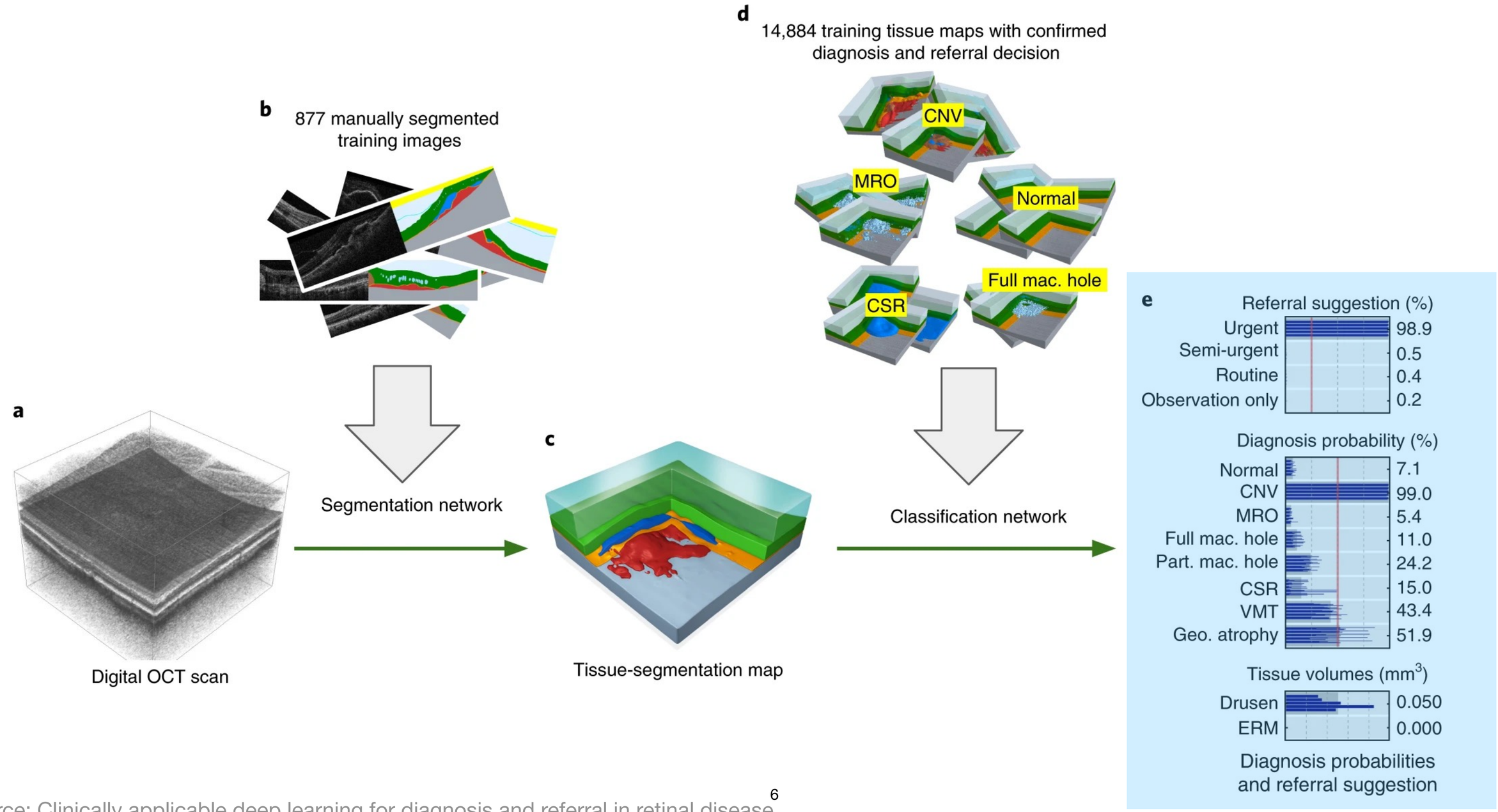
What is the probability that the user will click on one of these?

Probability examples and why we need to care



What is the probability that the current viewer will click on these?

Probability examples and why we need to care



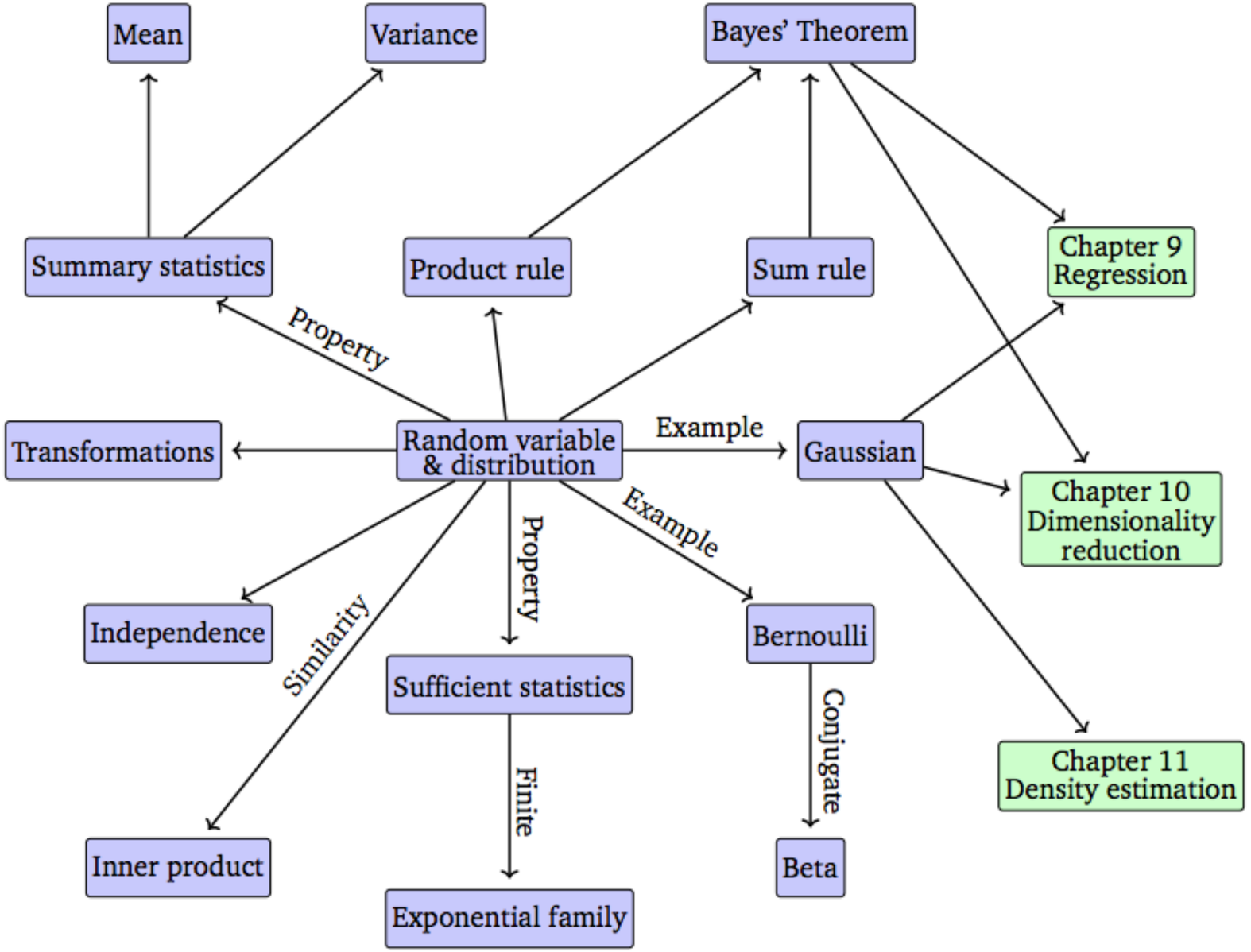
Source: Clinically applicable deep learning for diagnosis and referral in retinal disease

Probability examples and why we need to care

The key idea behind the probabilistic framework to machine learning is that learning can be thought of as inferring plausible models to explain observed data. A machine can use such models to make predictions about future data, and take decisions that are rational given these predictions. Uncertainty plays a fundamental part in all of this. Observed data can be consistent with many models, and therefore which model is appropriate, given the data, is uncertain. Similarly, predictions about future data and the future consequences of actions are uncertain. Probability theory provides a framework for modelling uncertainty.

Source: Probabilistic machine learning and artificial intelligence, Zoubin Ghahramani, Nature 2015

Overview


Mean
Variance
Bayes' Theorem
Summary statistics
Product rule
Sum rule
Chapter 9
Regression
Property
Transformations
Random variable
& distribution
Example
Gaussian
Chapter 10
Dimensionality
reduction
Example
Property
Independence
Similarity
Sufficient statistics
Bernoulli
Conjugate
Chapter 11
Density estimation
Finite
Inner product
Beta
Exponential family

An example

Outcome of a coin flip, O, is ***random***

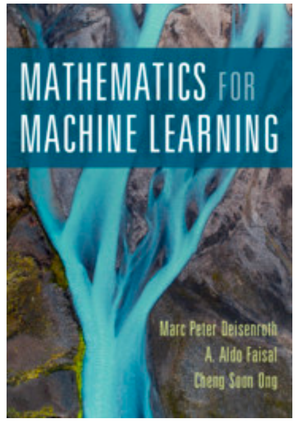
There are ***two*** possible outcomes: head (H) or tail (T)

Tail Head

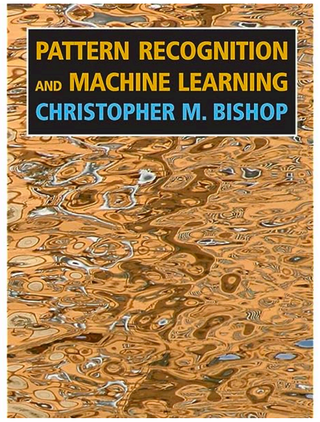
Questions we may want to ask:

- What is the probability of getting a head?
- What is the probability of getting a tail?
- Is it a fair coin?
- If we flip the coin *many many* times and we get one dollar for a head and zero for a tail, how much money will we make?
- What happens if instead we get 2 dollars per head and lose 25 dollars per tail?

One more example



I



S

ANU computing students take both *Introduction to Machine Learning (**I**)*

and *Statistical Machine Learning (**S**)*

Potential mark for each course: HD (**h**) or non-HD (**n**)

Variable: marks for both courses (**I, S**)

There are 4 possible outcomes: (I = h, S = h), (I = h, S = n), (I = n, S = h), (I = n, S = n)

Questions we may want to ask:

- What is the probability of students doing well in IML, aka getting an HD in IML?
- What is the probability of students not doing well in SML, getting a non-HD in SML?
- What is the probability of getting at least one HD?
- What is the probability of getting an HD in SML given an HD in IML?
- What is the probability of getting an non-HD in IML given an HD in SML?

Random variables - Discrete [1]

A *random* variable: possible values are outcomes of a random phenomenon.

There are two types of random variables, *discrete* and *continuous*.

Discrete random variable: outcome space is *discrete* [Head/Tail, HD/non-HD, or 0, 1, 2, 3, 4…]

The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values. It is also called the probability mass function (*pmf*).

We write P(X = *a*): the probability that the random variable X takes on the value *a*

- express the plausibility / belief about X = *a*
- $0 \leq P(X = a) \leq 1$
- Probabilities must sum to one: $\sum_{a \text{ in outcome space}} P(X = a) = 1$

Random variables - Discrete [2]

The joint probability distribution of multiple discrete random variables:

1. A list of probabilities associated with each of their possible values, or
2. A multi-dimensional array of these probabilities.

ANU computing students take both *Introduction to Machine Learning (**I**)* and *Statistical Machine Learning (**S**)*

Potential mark for each course: HD (**h**) or non-HD (**n**)

Variable: marks for both courses (**I, S**)

There are 4 possible outcomes: (I = h, S = h), (I = h, S = n), (I = n, S = h), (I = n, S = n)

1

P(I = h, S = h)
P(I = h, S = n)
P(I = n, S = h)
P(I = n, S = n)

2

	S = h	S = n
I = h	P(I = h, S = h)	P(I = h, S = n)
I = n	P(I = n, S = h)	P(I = n, S = n)

Random variables - Discrete - a numerical example

$0 \leq$ Probabilities ≤ 1, and sum of all probabilities = 1

	S = h	S = n
I = h	P(I = h, S = h) = 0.2	P(I = h, S = n) = 0.15
I = n	P(I = n, S = h) = 0.05	P(I = n, S = n) = 0.6

Random variables - Discrete - a numerical example

Marginal probability = probability of one or more variables ***irrespective*** of the value of the other variables
= sum of a row or column or dimension

	S = h	S = n
I = h	P(I = h, S = h) = 0.2	P(I = h, S = n) = 0.15
I = n	P(I = n, S = h) = 0.05	P(I = n, S = n) = 0.6

Marginal here means summing out or integrate out all other variables!

- What is the probability of students doing well in IML, aka getting an HD in IML?

- What is the probability of students not doing well in SML, getting a non-HD in SML?

Random variables - Discrete - a numerical example

Conditional probability = probability of one or more variables ***given*** of the value of the other variables
= fraction of a cell in a row or column or dimension

	S = h	S = n
I = h	P(I = h, S = h) = 0.2	P(I = h, S = n) = 0.15
I = n	P(I = n, S = h) = 0.05	P(I = n, S = n) = 0.6

- What is the probability of getting an HD in SML given an HD in IML?

- What is the probability of getting an non-HD in IML given an HD in SML?

Two fundamental rules - Sum and Product

Terminology:

- P(X, Y) is the *joint* distribution of two random variables X and Y
- P(X), P(Y) are the *marginal* distributions
- P(Y | X) is the *conditional* distribution of Y given X
- P(X | Y) is the *conditional* distribution of X given Y

Sum rule

$$P(X) = \sum_{y \in \mathcal{Y}} P(X, Y = y)$$

where $\mathcal{Y}$ is the outcome space of Y. Also called marginalisation property.

Product rule

$$P(X, Y) = P(X \mid Y)P(Y) = P(Y \mid X)P(X)$$

Check worked example in previous slides

Bayes' rule - inverse probability

Product rule: $$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

We can rewrite: $$P(X|Y) = \frac{P(X,Y)}{P(Y)} = \frac{P(X)\,P(Y|X)}{P(Y)}$$ or $P(Y|X) = \frac{P(X,Y)}{P(X)} = \frac{P(Y)\,P(X|Y)}{P(X)}$

This is called **Bayes' rule** or Bayes' theorem.

Posterior
belief about X after knowing Y

$=$

$$\frac{\textbf{Prior}\ \text{prior belief about X} \times \textbf{Likelihood}\ \text{likelihood of X given Y}}{\textbf{Marginal likelihood}\ \text{or evidence}}$$

likelihood averaged over all potential X

Machine learning model + Bayes = *probabilistic/Bayesian machine learning [SML, S1 2024]*

X is the model or model parameters [*unknown*], Y = data [*observed*]

Learning = *computing* the posterior; model/hyperparameter selection using *evidence*;

Prediction = *summing* over all plausible models/model parameters

An inference example

My neighbor has two children. Assuming that the gender of a child is like a coin flip, it is most likely, a priori, that my neighbor has one boy and one girl, with probability 1/2. The other possibilities—two boys or two girls—have probabilities 1/4 and 1/4.

a. Suppose I ask him whether he has *any* boys, and he says *yes*. What is the probability that one child is a girl?

b. Suppose instead that I happen to see one of his children run by, and it is a boy. What is the probability that the other child is a girl?

Source: Tom Minka and PML book (Murphy)

Random variables - Continuous [1]

Many physical measurements or parameters in ML can take any value in a continuous range.

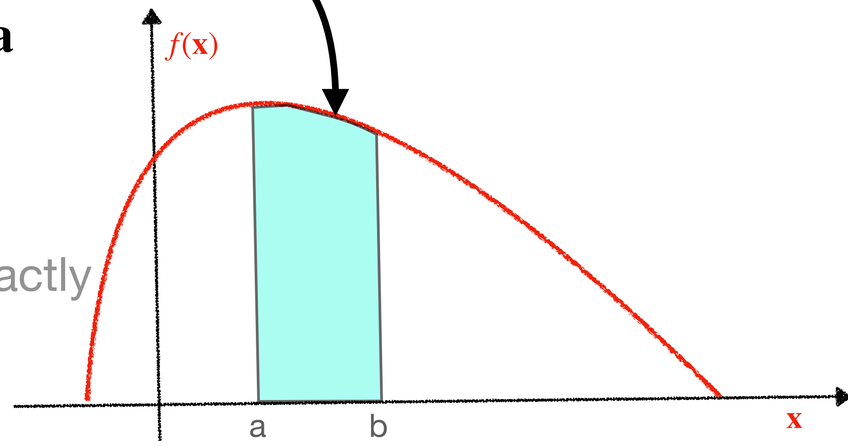
For D-dimensional continuous random variables **X** [each dimension is a random variable], we associate **X** with a probability density function (pdf):

$$f : \mathbb{R}^D \to \mathbb{R} \qquad \forall \mathbf{x} \in \mathbb{R}^D : f(\mathbf{x}) \geq 0 \qquad \text{and} \qquad \int_{\mathbb{R}^D} f(\mathbf{x}) \mathrm{d}\mathbf{x} = 1$$

Reminder: in the discrete case [slide 11], the pmf satisfies: $0 \leq P(X = a) \leq 1$ and $\sum_a P(X = a) = 1$.

We say **X** is distributed according to $f(\mathbf{x})$ if $P(\mathbf{a} \leq \mathbf{X} \leq \mathbf{b}) = \int_{\mathbf{a}}^{\mathbf{b}} f(\mathbf{x}) \mathrm{d}\mathbf{x}$

Note: $P(\mathbf{X} = \mathbf{a}) = \int_{\mathbf{a}}^{\mathbf{a}} f(\mathbf{x}) \mathrm{d}\mathbf{x} = 0$



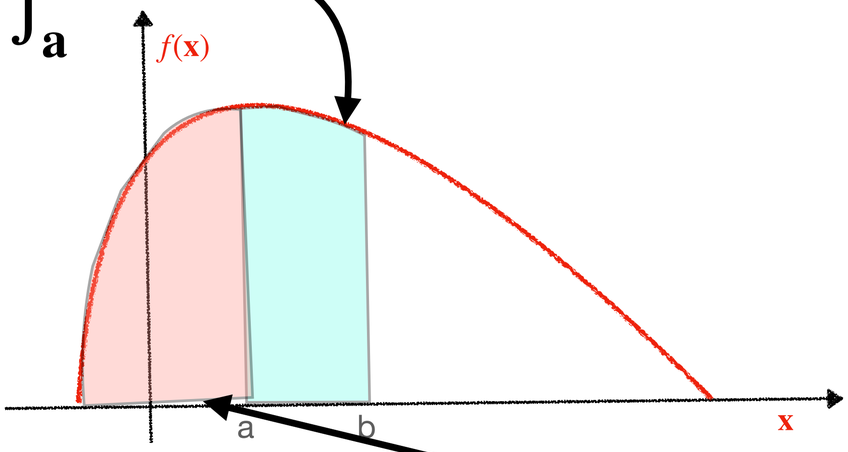
Hand-wavy: we can measure **X** between **a** and **a** + δ, but can never say **X** = **a** exactly

We often write p(**x**) instead of f(**x**). For continuous r.v., p(**x**) is not probability!

Random variables - Continuous [2]

We say **X** is distributed according to $f(\mathbf{x})$ if $P(\mathbf{a} \leq \mathbf{X} \leq \mathbf{b}) = \int_{\mathbf{a}}^{\mathbf{b}} f(\mathbf{x})\mathrm{d}\mathbf{x}$

Area under f(**x**) between **a** and **b**



Area under f(**z**) on the left of **x**

We can also find $F_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f(\mathbf{z})\mathrm{d}\mathbf{z}$

This is called the cumulative distribution function (cdf).

Note that: $P(\mathbf{a} \leq \mathbf{X} \leq \mathbf{b}) = \int_{\mathbf{a}}^{\mathbf{b}} f(\mathbf{x})\mathrm{d}\mathbf{x} = \int_{-\infty}^{\mathbf{b}} f(\mathbf{z})\mathrm{d}\mathbf{z} - \int_{-\infty}^{\mathbf{a}} f(\mathbf{z})\mathrm{d}\mathbf{z} = \mathbf{F}(\mathbf{b}) - \mathbf{F}(\mathbf{a})$

and $f(\mathbf{x}) = \frac{d}{d\mathbf{x}} F(\mathbf{x})$

Example: Find $f(x)$, given $F(x) = \begin{cases} 0 & \text{if } x < 1 \\ 2(x-1) & \text{if } 1 \leq x \leq 1.5 \\ 1 & \text{if } x > 1.5 \end{cases}$

Distributions: discrete vs continuous

Discrete:

Probability mass function (pmf): $\forall a : 0 \leq P(X = a) \leq 1$ and $\sum_a P(X = a) = 1$

Sum rule $P(X) = \sum_{y \in \mathcal{Y}} P(X, Y = y)$ **Product rule** $P(X, Y) = P(X \mid Y)P(Y) = P(Y \mid X)P(X)$

Continuous:

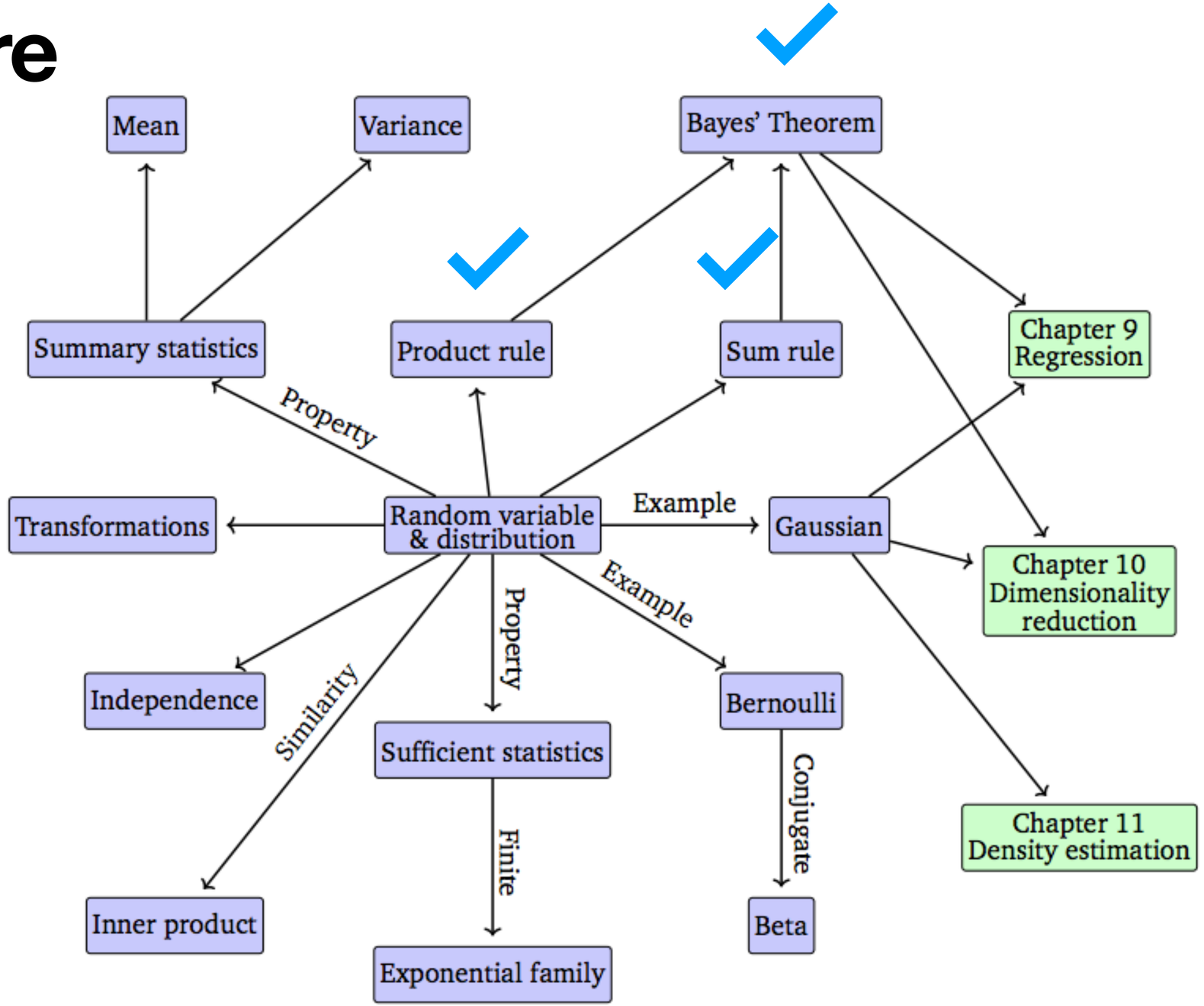
Probability density function (pdf): $\forall \mathbf{x} \in \mathbb{R}^D : f(\mathbf{x}) \geq 0$ and $\int_{\mathbb{R}^D} f(\mathbf{x})\mathrm{d}\mathbf{x} = 1$

Cumulative density function (cdf): $F_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f(\mathbf{z})\mathrm{d}\mathbf{z}$

Sum rule $f(\mathbf{x}) = \int_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\mathrm{d}\mathbf{y}$ **Product rule** $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x} \mid \mathbf{y})f(\mathbf{y}) = f(\mathbf{y} \mid \mathbf{x})f(\mathbf{x})$

Note: f(x) can be larger than 1. Example: the *uniform* example in the last slide

Big picture


Mean
Variance
Bayes' Theorem
Summary statistics
Product rule
Sum rule
Chapter 9
Regression
Property
Transformations
Random variable
& distribution
Example
Gaussian
Chapter 10
Dimensionality
reduction
Example
Property
Independence
Similarity
Sufficient statistics
Bernoulli
Conjugate
Chapter 11
Density estimation
Finite
Inner product
Beta
Exponential family

Summary statistics - expected value

Example:

Tail

Head

If we flip the coin *many many* times and we get \$3 for a head and \$1 for a tail, how much money will we make?

We can write: expected value = \$3 × P(H) + \$1 × P(T) = \$3 × p + \$1 × (1-p) = \$(2p + 1)

Definition The expected value of a function g of a univariate random variable $X \sim p(x)$:

$$\mathbb{E}_X[g(x)] = \begin{cases} \sum_{x \in \mathcal{X}} g(x)p(x) & \text{if } X \text{ is discrete} \\ \int_{\mathcal{X}} g(x)p(x)\mathrm{d}x & \text{if } X \text{ is continuous} \end{cases}$$

For multivariate random variables, the expected value is computed *element-wise*:

$$\mathbb{E}_{\mathbf{X}}[g(\mathbf{x})] = \begin{bmatrix} \mathbb{E}_{X_1}[g(x_1)] \\ \vdots \\ \mathbb{E}_{X_D}[g(x_D)] \end{bmatrix}$$

Example: A bookie advertises a lottery with a winning chance of 1 in 500,000. The winning price is \$800,000. To enter, you have to buy a \$2 ticket. Should you play?

Summary statistics - mean

When g(x) = x, we call the expected value the mean of the distribution

$$\mathbb{E}_X[x] = \begin{cases} \sum_{x \in \mathcal{X}} xp(x) & \text{if } X \text{ is discrete} \\ \int_{\mathcal{X}} xp(x)\mathrm{d}x & \text{if } X \text{ is continuous} \end{cases}$$

Example: Compute the mean of a Poisson distribution with parameter λ, given its pmf,

$$p(k) = P(X = k) = \frac{\lambda^k \exp(-\lambda)}{k!}$$

Summary statistics - covariance

We may wish to determine the *joint variability* of two variables, by computing the covariance:

$$\mathrm{Cov}_{X,Y}[x, y] = \mathbb{E}_{X,Y}[(x - \mu_x)(y - \mu_y)]$$

where μ_x and μ_y the the mean of X and Y respectively.

Notes:

1. $\mathrm{Cov}_{X,Y}[x, y] = \mathbb{E}_{X,Y}[xy] - \mu_x\mu_y$
2. For univariate distributions, $\mathrm{Cov}_{X,X}[x, x] = \mathbb{E}_X[x^2] - \mu_x^2$ is called the variance, and its square root is called the standard deviation
3. For multivariate distributions, $\mathrm{Cov}_{\mathbf{X},\mathbf{Y}}[\mathbf{x}, \mathbf{y}] = \mathbb{E}_{\mathbf{X},\mathbf{Y}}[\mathbf{xy}] - \mu_\mathbf{x}\mu_\mathbf{y}^\intercal = \mathrm{Cov}_{\mathbf{Y},\mathbf{X}}[\mathbf{y}, \mathbf{x}]^\intercal$

Example Compute the variance of a continuous uniform dist. with support between [a, b]

Summary statistics - variance and correlation

When $\mathbf{y} = \mathbf{x}, \mathrm{Cov}_{\mathbf{X}}[\mathbf{x}, \mathbf{x}]$ is called the variance, or covariance matrix of $\mathbf{X}$:

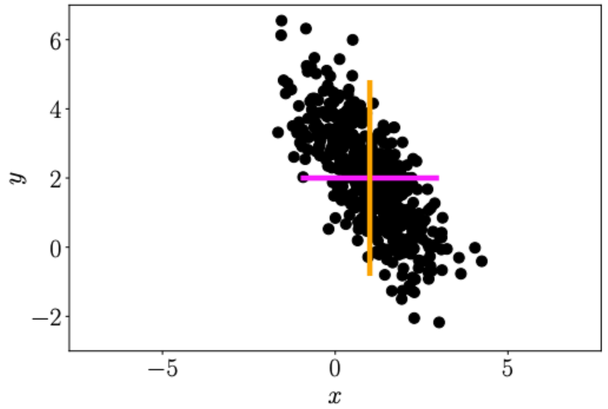
$$\mathbb{V}_{\mathbf{X}}[\mathbf{x}] = \mathrm{Cov}_{\mathbf{X}}[\mathbf{x}, \mathbf{x}] = \mathbb{E}_{\mathbf{X}}[(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})^{\intercal}] = \begin{bmatrix} \mathrm{Cov}[x_1, x_1] & \mathrm{Cov}[x_1, x_2] & \cdots & \mathrm{Cov}[x_1, x_D] \\ \mathrm{Cov}[x_2, x_1] & \mathrm{Cov}[x_2, x_2] & \cdots & \mathrm{Cov}[x_2, x_D] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[x_D, x_1] & \mathrm{Cov}[x_D, x_2] & \cdots & \mathrm{Cov}[x_D, x_D] \end{bmatrix}$$

Cross-covariance

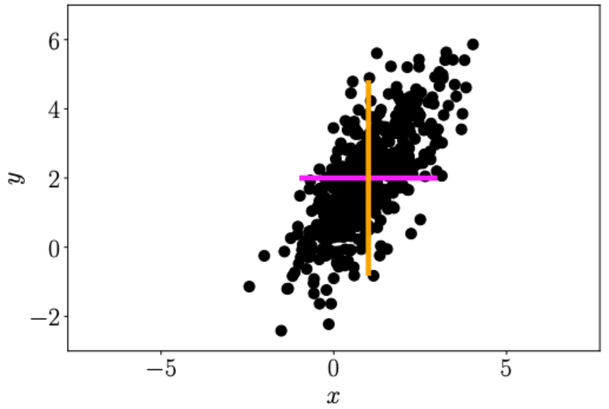
Marginal variance

We can normalise the covariance by the variances, this is called the correlation:

$$\mathrm{corr}[x, y] = \frac{\mathrm{Cov}[x, y]}{\sqrt{\mathbb{V}[x]\mathbb{V}[y]}} \in [-1, 1]$$

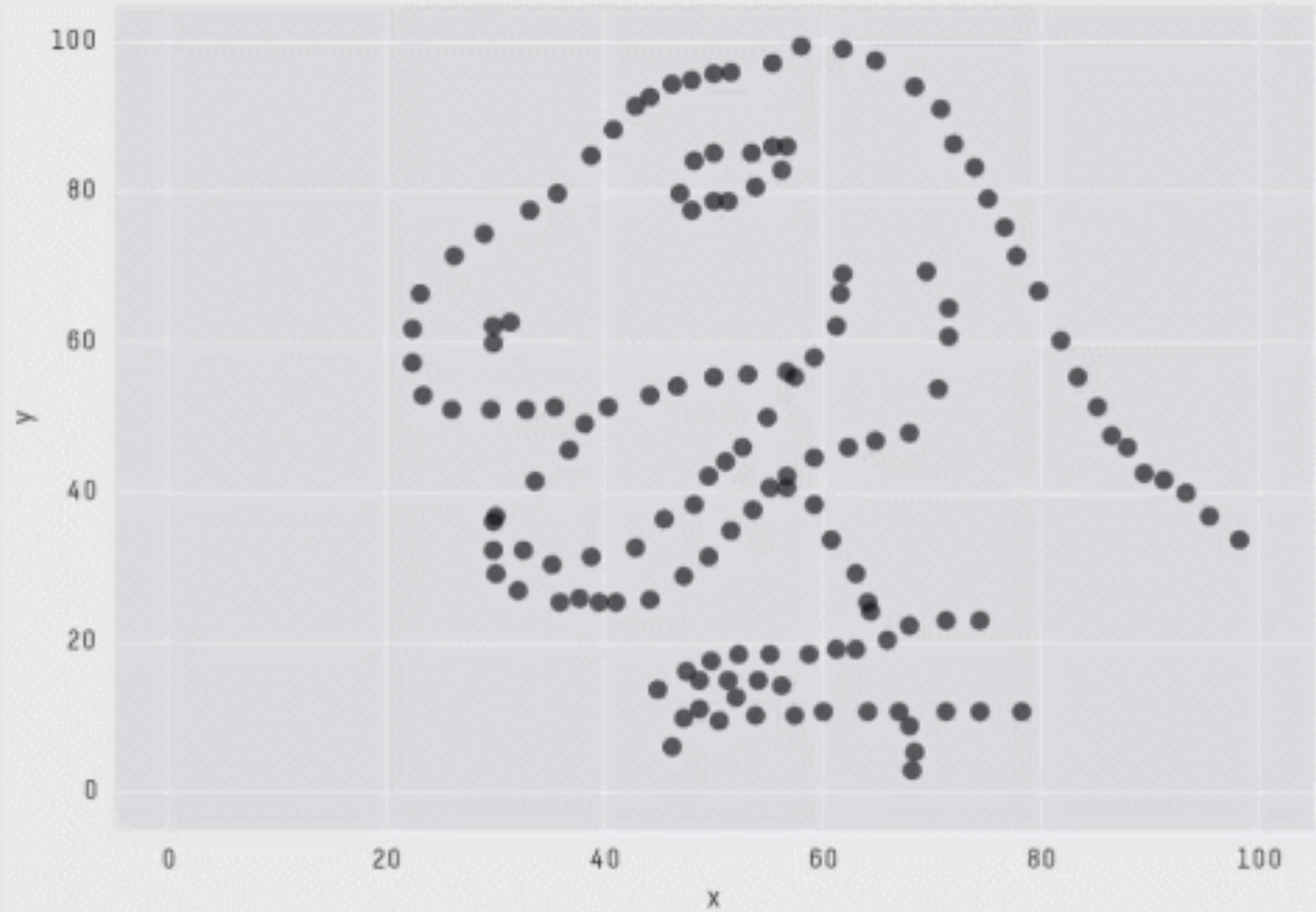


(a) x and y are negatively correlated.



(b) x and y are positively correlated.

Summary statistics, correlations and a cautionary note



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Your data have more than just summary statistics!

Source: Datasaurus Dozen

Empirical means and covariances

In many statistical applications such as machine learning, we often have access only to the finite number of data samples or observations, not the full underlying distributions.

We can estimate the mean and covariance using:

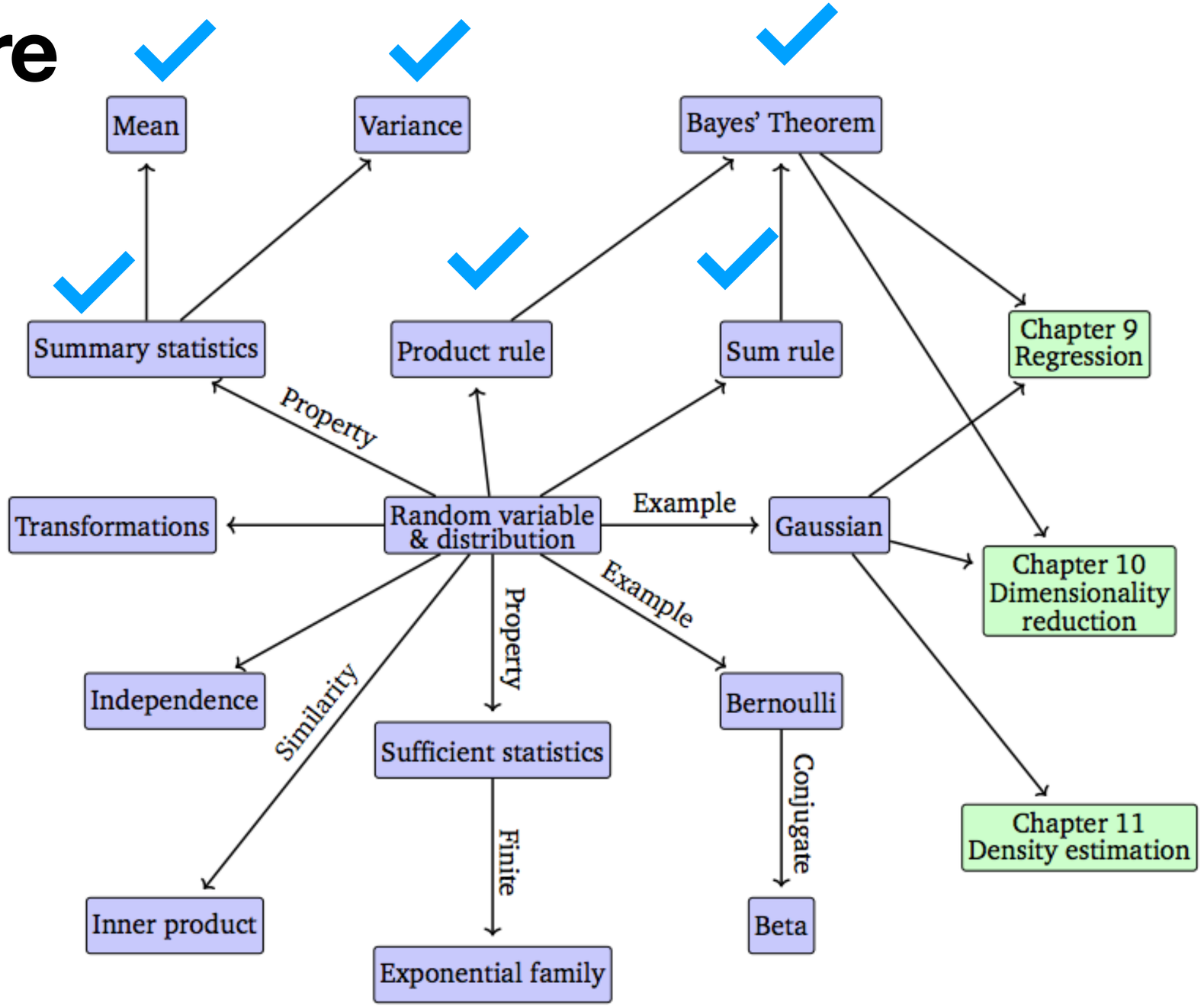
$$\bar{\mathbf{x}} := \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \qquad \Sigma := \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^\intercal$$

Empirical mean Empirical covariance

Questions

1. What is the computational complexity of these? Can we do in a single pass?
2. Can we change these to handle streaming data?
3. The empirical covariance above is a biased estimate. Why and how can we fix this?

Big picture


Mean
Variance
Bayes' Theorem
Summary statistics
Product rule
Sum rule
Chapter 9
Regression
Property
Transformations
Random variable
& distribution
Example
Gaussian
Chapter 10
Dimensionality
reduction
Example
Property
Independence
Similarity
Sufficient statistics
Bernoulli
Conjugate
Chapter 11
Density estimation
Finite
Inner product
Beta
Exponential family

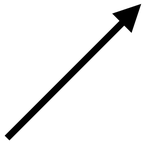
Sums and transformations of random variables

Means and covariances of sum of random variables, $X + Y$ or $X - Y$

$$\mathbb{E}[\mathbf{x} + \mathbf{y}] = \mathbb{E}[\mathbf{x}] + \mathbb{E}[\mathbf{y}]$$
$$\mathbb{E}[\mathbf{x} - \mathbf{y}] = \mathbb{E}[\mathbf{x}] - \mathbb{E}[\mathbf{y}]$$
$$\mathbb{V}[\mathbf{x} + \mathbf{y}] = \mathbb{V}[\mathbf{x}] + \mathbb{V}[\mathbf{y}] + \mathrm{Cov}[\mathbf{x}, \mathbf{y}] + \mathrm{Cov}[\mathbf{y}, \mathbf{x}]$$
$$\mathbb{V}[\mathbf{x} - \mathbf{y}] = \mathbb{V}[\mathbf{x}] + \mathbb{V}[\mathbf{y}] - \mathrm{Cov}[\mathbf{x}, \mathbf{y}] - \mathrm{Cov}[\mathbf{y}, \mathbf{x}]$$

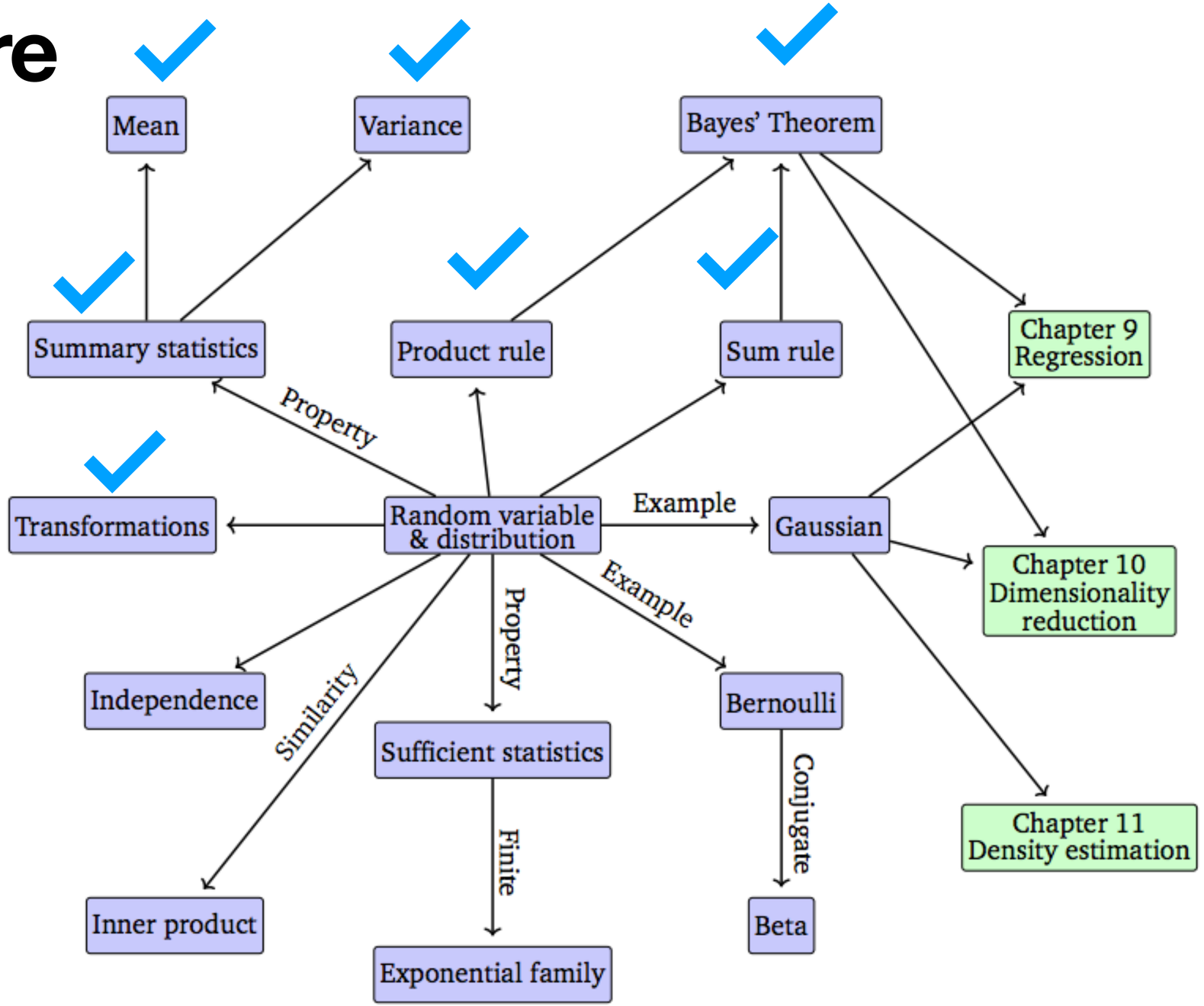
Means and covariances of affine transformation, $Y = \mathbf{A}X + \mathbf{b}$

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\mathbb{E}[\mathbf{x}] + \mathbf{b}$$
$$\mathbb{V}[\mathbf{y}] = \mathbb{V}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\mathbb{V}[\mathbf{x}]\mathbf{A}^\intercal$$
$$\mathrm{Cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}[\mathbf{x}(\mathbf{A}\mathbf{x} + \mathbf{b})^\intercal] - \mathbb{E}[\mathbf{x}]\mathbb{E}[(\mathbf{A}\mathbf{x} + \mathbf{b})]^\intercal = \mathbb{V}[\mathbf{x}]\mathbf{A}^\intercal$$



Crucial for Gaussian manipulation: Linear regression [Week 7], dynamical systems [Kalman filters], Gaussian processes [SML S1 2024]

Big picture


Mean
Variance
Bayes' Theorem
Summary statistics
Product rule
Sum rule
Chapter 9
Regression
Property
Transformations
Random variable
& distribution
Example
Gaussian
Chapter 10
Dimensionality
reduction
Example
Property
Similarity
Independence
Sufficient statistics
Bernoulli
Conjugate
Chapter 11
Density estimation
Finite
Inner product
Beta
Exponential family

Statistical independence

Two random variables X, Y are statistically independent *if and only if* $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$

This implies:

$$p(\mathbf{y} \mid \mathbf{x}) = p(\mathbf{y})$$
$$p(\mathbf{x} \mid \mathbf{y}) = p(\mathbf{x})$$

Knowing x does not add any additional information about y, and vice versa

$$\mathbb{V}[\mathbf{x} + \mathbf{y}] = \mathbb{V}[\mathbf{x}] + \mathbb{V}[\mathbf{y}]$$
$$\mathrm{Cov}[\mathbf{x}, \mathbf{y}] = \mathbf{0}$$

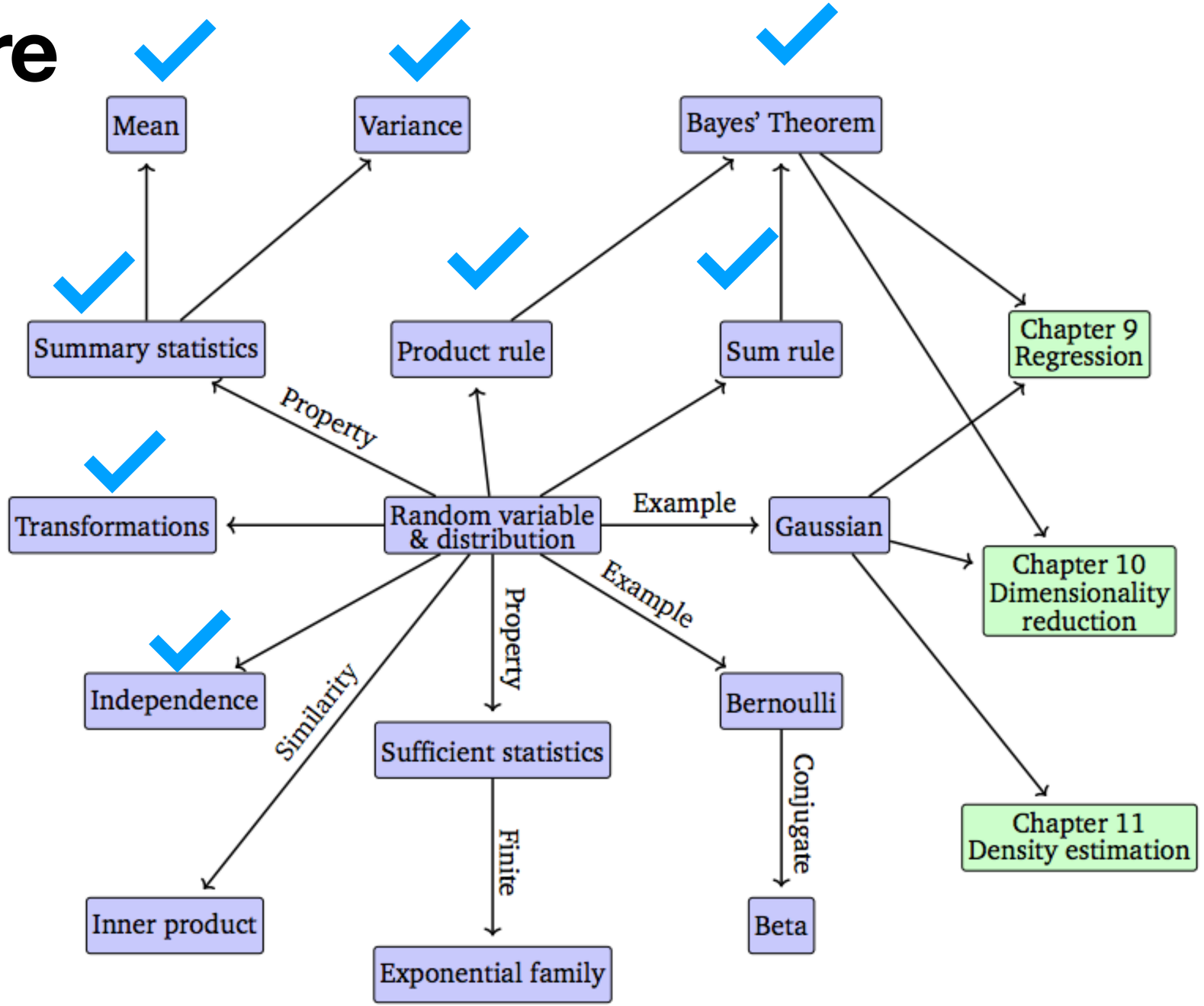
Reverse is not true, as covariance measures linear dependence
Remember the data dinosaur on slide 27!

Two random variables X, Y are conditionally independent *given Z if and only if*

$p(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{y} \mid \mathbf{z})$, aka *given Z*, knowing X does not give any new information about Y and vice versa.

These properties are heavily used in information-theoretic active learning, unsupervised learning, causal estimation, and inference and learning in graphical models.

Big picture


Mean
Variance
Bayes' Theorem
Summary statistics
Product rule
Sum rule
Chapter 9
Regression
Property
Transformations
Random variable
& distribution
Example
Gaussian
Chapter 10
Dimensionality
reduction
Example
Property
Independence
Similarity
Sufficient statistics
Bernoulli
Conjugate
Chapter 11
Density estimation
Finite
Inner product
Beta
Exponential family

Some named distributions

Distribution	Example
Bernoulli	Outcome of a coin toss
Binomial	Outcome of multiple coin tosses
Geometric	Number of coin tosses until first head
Poisson	Number of events within a given time interval
Normal (Gaussian)	Daily financial returns, temperature
Exponential	Time between events, e.g. time until a radioactive particle decays

Bernoulli and Binomial distributions

Bernoulli: outcome of a coin toss - head or tail. Probability of head = p, tail = 1 - p

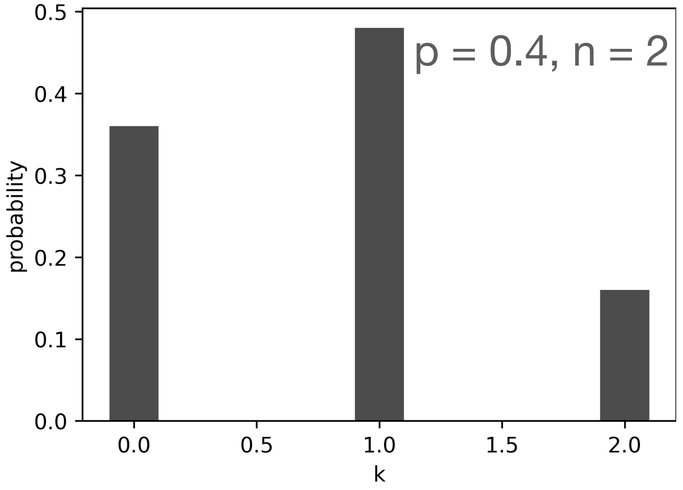
Binomials: number of heads in n coin tosses

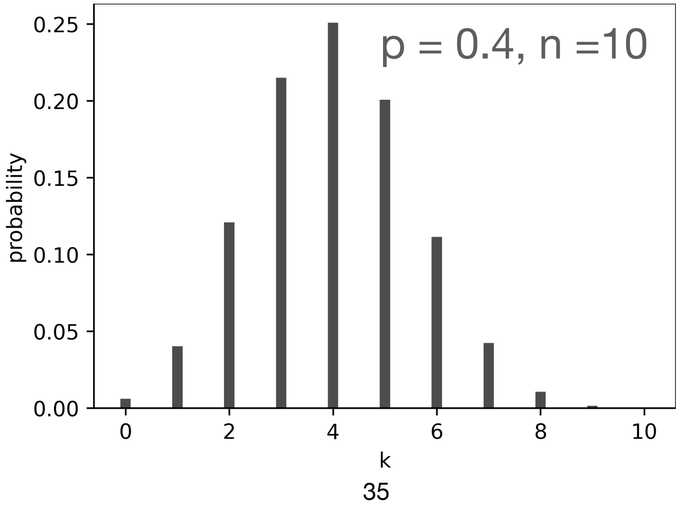
Example: n = 2, probability of 0, 1, 2 heads are:

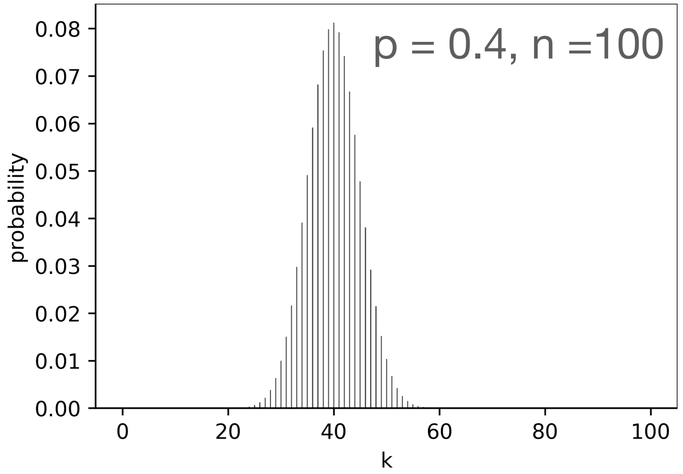
$p_{0,2} =$ $p_{1,2} =$ $p_{2,2} =$

Probability of exactly k heads in n tosses:

$$p_{k,n} = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{with} \quad \binom{n}{k} = \frac{n!}{k!(n-k)!} \quad \text{and} \quad 0! = 1$$







Gaussian distributions - univariate

A continuous, real-valued, univariate **Gaussian** or **normal** random variable has the following probability density function

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

mean = μ, variance = σ^2, standard deviation = σ

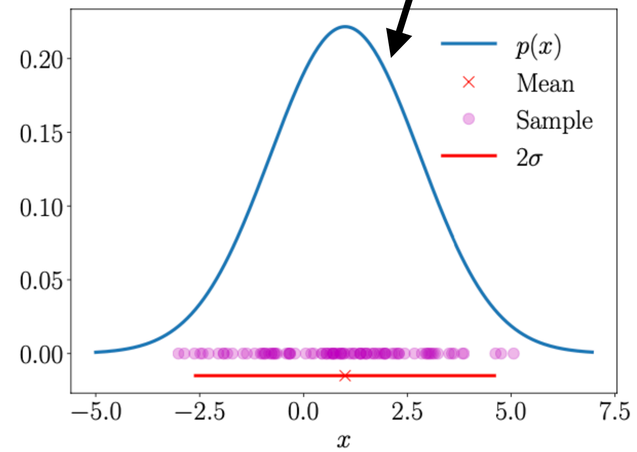
We often write: $p(x \mid \mu, \sigma^2) = \mathcal{N}(x; \mu, \sigma^2)$ or $x \sim \mathcal{N}(\mu, \sigma^2)$

Standard normal distribution: $\mu = 0$ and $\sigma = 1$

Cumulative probability distribution = probability to the left of x

$$\text{cdf}(x) = \int_{-\infty}^{x} p(z)\text{d}z = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

bell shaped curve



(a) Univariate (one-dimensional) Gaussian; The red cross shows the mean and the red line shows the extent of the variance.

Gaussian distributions - example 1

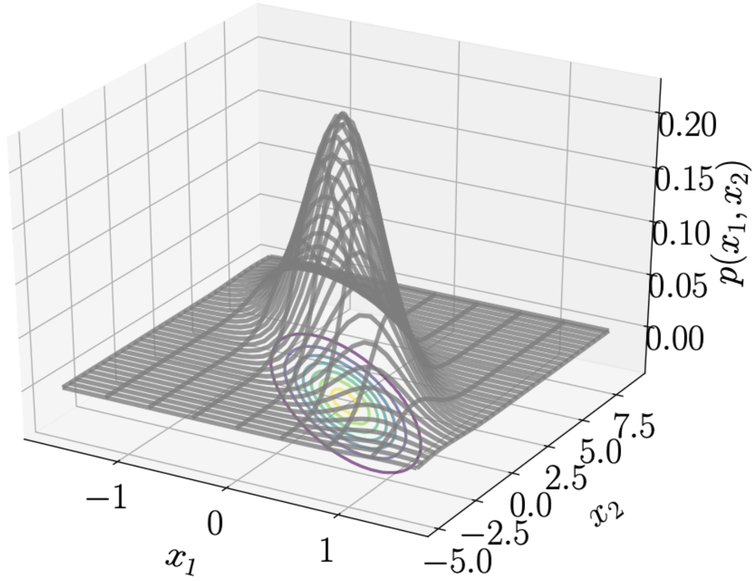
(a) if $p(x) = \exp(-ax^2 + bx + c)$ is a probability density, what are the mean and variance?

(b) find $\int_{-\infty}^{\infty} \exp(-0.5x^2)\mathrm{d}x$

Gaussian distributions - multivariate

A D-dimensional multivariate **Gaussian** or **normal** distribution is characterised by a *mean* vector μ and a *covariance matrix* Σ, with

pdf: $p(\mathbf{x} \mid \mu, \Sigma) = (2\pi)^{-\frac{D}{2}} \mid \Sigma \mid^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right)$

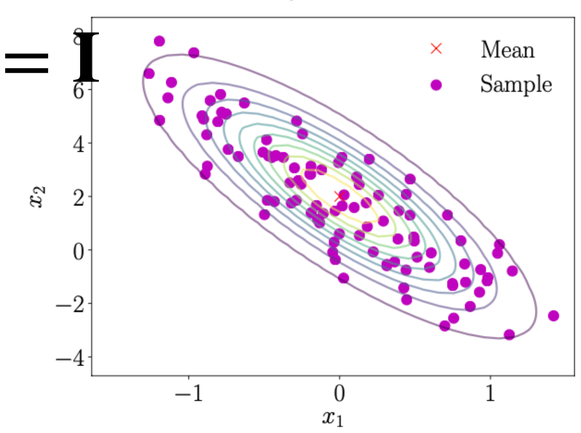


We often write: $p(\mathbf{x} \mid \mu, \Sigma) = \mathcal{N}(\mathbf{x}; \mu, \Sigma)$ or $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$

Standard multivariate **Gaussian** or **normal** distribution: $\mu = \mathbf{0}$ and $\Sigma = \mathbf{I}$

When Σ is a diagonal matrix, $p(\mathbf{x} \mid \mu, \Sigma) = \prod_{d=1}^{D} p(x_d \mid \mu_d, \Sigma_{d,d})$

Diagonal Gaussian, dimensions are independent



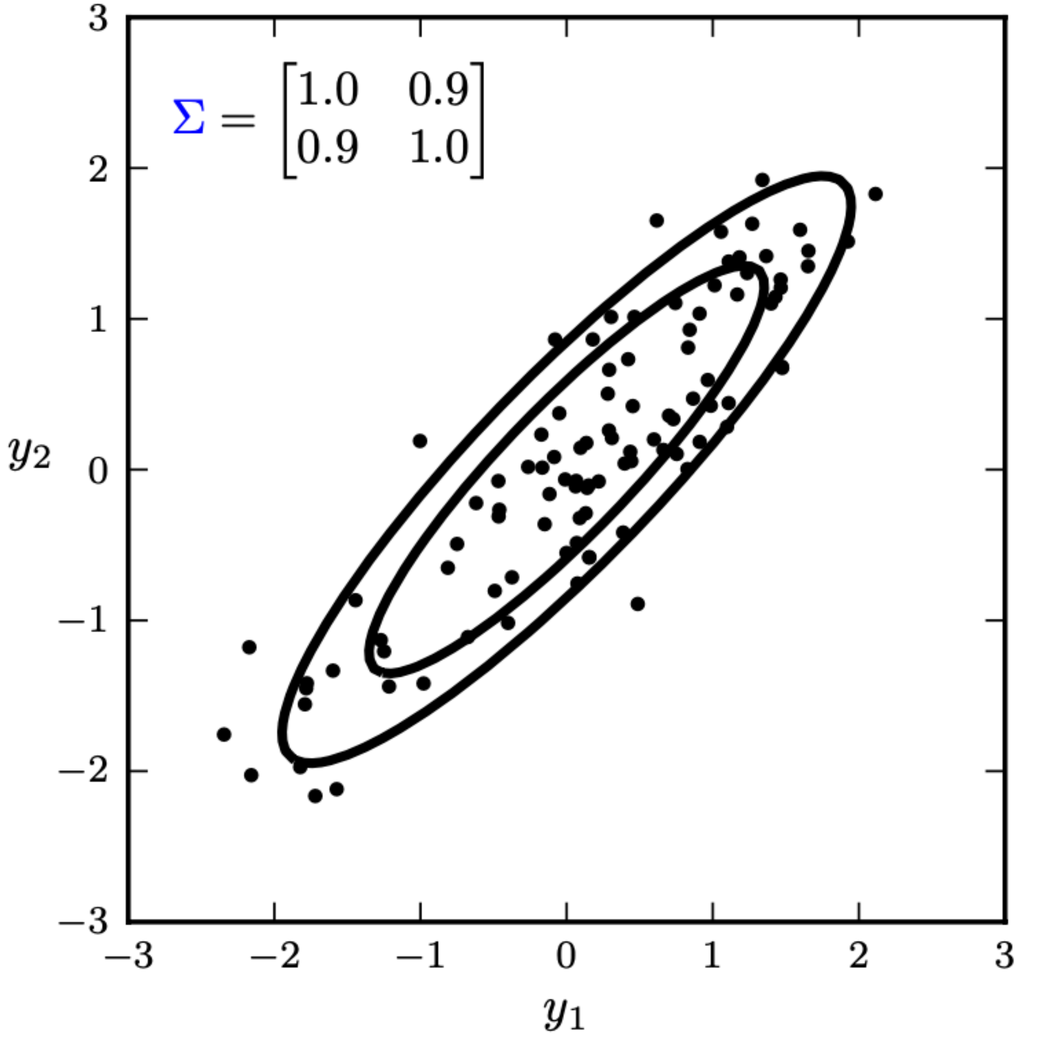
(b) Multivariate (two-dimensional) Gaussian, viewed from top. The red cross shows the mean and the colored lines show the contour lines of the density.

Alternative parameterisation, $\eta_1 = \Sigma^{-1}\mu$ and $\eta_2 = \Sigma^{-1}$ [precision]

Easier for multiplication/division and identifying conditional independence

Gaussian distributions - multivariate viz

$$p(\mathbf{y}|\Sigma) \propto \exp(-\tfrac{1}{2}\mathbf{y}^T\Sigma^{-1}\mathbf{y})$$



Gaussian distributions - marginal and conditional

Let $\mathbf{X}$ and $\mathbf{Y}$ be two multivariate random variables that may have different dimensions.

The joint pdf is $p(\mathbf{x}, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}; \begin{bmatrix} \mu_{\mathbf{x}} \\ \mu_{\mathbf{y}} \end{bmatrix}, \begin{bmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}} \end{bmatrix} \right)$

Cross covariance

Marginal covariance

Marginal and conditional distributions are Gaussian!

Marginal

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}; \mu_{\mathbf{x}}, \Sigma_{\mathbf{xx}}\right)$$

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y}; \mu_{\mathbf{y}}, \Sigma_{\mathbf{yy}}\right)$$

Conditional

$$p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{x}; \mu_{\mathbf{x}|\mathbf{y}}, \Sigma_{\mathbf{x}|\mathbf{y}}\right)$$

$$\mu_{\mathbf{x}|\mathbf{y}} = \mu_{\mathbf{x}} + \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1}(\mathbf{y} - \mu_{\mathbf{y}})$$

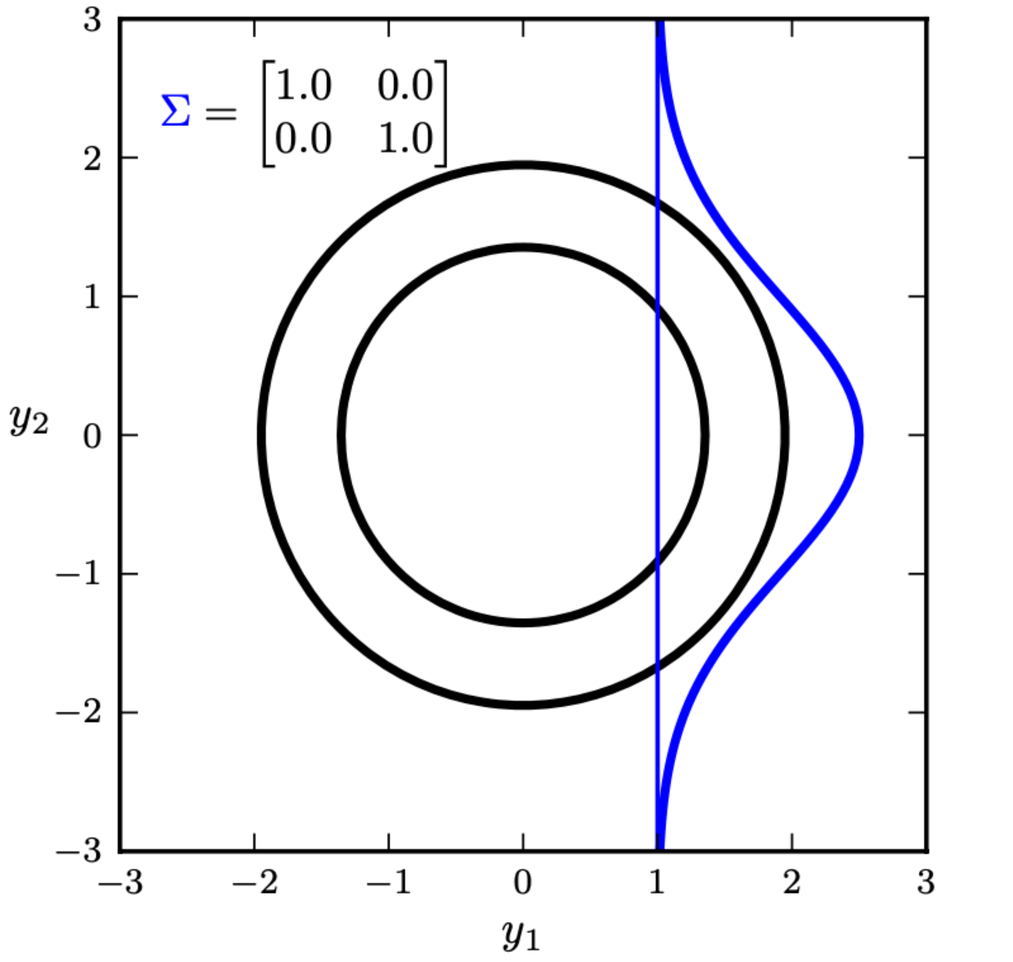
$$\Sigma_{\mathbf{x}|\mathbf{y}} = \Sigma_{\mathbf{xx}} - \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{yx}}$$

Other interesting properties:

- Sum of independent Gaussian random variables is Gaussian
- Product of Gaussians is Gaussian [see next worked example]
- Convolution of two Gaussians is Gaussian
- Linearly transformed Gaussian random variable is Gaussian

Gaussian distributions - conditional viz

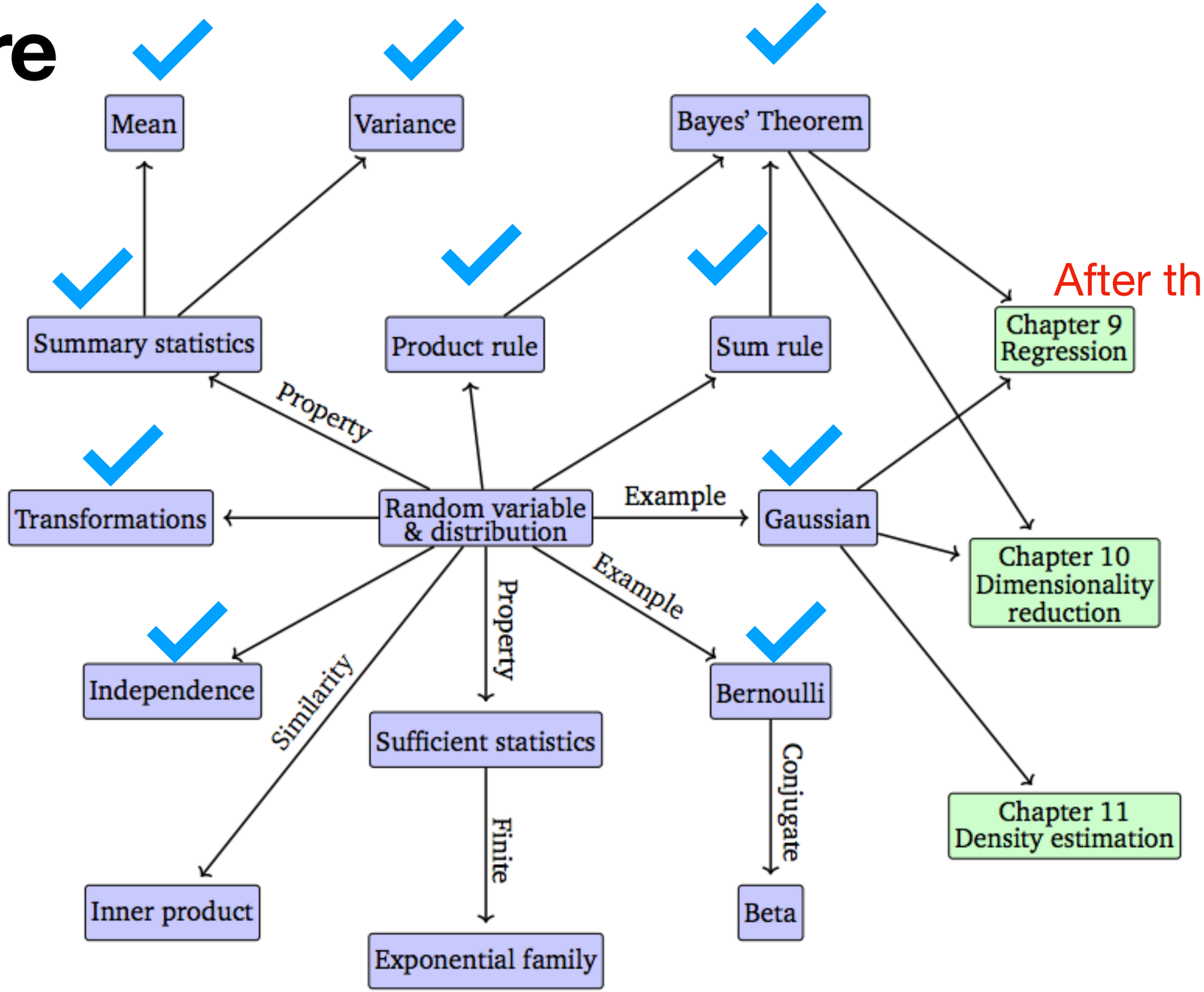
$$p(y_2|y_1, \Sigma) \propto \exp(-\tfrac{1}{2}(y_2 - \mu_2)^T {\Sigma_2}^{-1}(y_2 - \mu_2))$$



Gaussian distributions - example 2

Given $p_1(x) = \mathcal{N}(x; m_1, \sigma_1^2)$ and $p_2(x) = \mathcal{N}(x; m_2, \sigma_2^2)$, show that $p_1 p_2$ is also a Normal dist.

Big picture



After the break