



UNIVERSITY OF
OXFORD

The Exciting Guide To Probability Distributions – Part 1

Jamie Frost – v1.1

Overview

▶ Part 1

- ▶ The basics...
- ▶ Bernoulli
- ▶ Binomial
- ▶ Multinomial
- ▶ Geometric
- ▶ Gaussian
- ▶ Poisson

▶ Part 2

- ▶ Dirichlet
- ▶ Beta
- ▶ Conjugate Priors
- ▶ Gamma





The basics...

- ▶ **Random Variable:** The name of an experiment with a probabilistic outcome.
- ▶ **Probability Distribution:** For some random variable Y , specifies the prob $\Pr(Y=y_i)$ that Y will take value y_i (for all y_i).
- ▶ For convenience, we can write $\Pr(Y=y_i)$ as $P(y_i)$. *(but remember this is just a convenience, the input of the probability function is an **event**, not a value)*
- ▶ The **support** of a probability distribution can be loosely thought of as the closure of the set of possible values of a random variable having that distribution.

$$\Pr(X=x) = f: X \rightarrow (0, 1]$$

By definition the support doesn't contain values that map to a probability of 0, i.e. we don't give a monkeys about 'impossible events'.

The basics...



- ▶ Just like any old function, probability distributions can be **discrete** (i.e. although not necessarily finite), or **continuous** (e.g. picking a real number in the range $[0, 100]$).
- ▶ What distinguishes a probability distribution from a normal function is that the probabilities must sum to 1*.
- ▶ If we have a distribution D , then $x \sim D$ means to 'draw' a value from the distribution. For example, if we used $Dice_n$ to represent an n-sided dice, then $x \sim Dice_n$ means we have rolled the dice, and got the value x . **The drawn value is an element of the support**, and picked randomly accordingly to the associated probabilities.
- ▶ A **sample** is merely a number of draws from the (same) distribution. It is useful for estimating the underlying parameters of a distribution.

* At least in the discrete case. We'll see later what the requirement is for a continuous probability distribution.



The basics...

- ▶ The **expected value** $E[X]$ of a random variable X is (to be technical) its integral with respect to its underlying probability measure.
- ▶ In Layman's terms, it's the **average value** of X . For **discrete** distributions we can calculate it by summing over the domain of X , each outcome weighted by its probability:

$$E[X] = \sum_i x_i f(x_i)$$

- ▶ For a fair 3-sided dice this would be $(1/3)*1 + (1/3)*2 + (1/3)*3 = 2$.
- ▶ We'll revisit the mean for continuous distributions later.
- ▶ We express the mean as the symbol μ :

$$\mu = E[X]$$

The basics...

- ▶ Notice that $E[X]$ is dependent on the underlying distribution for X being **known**. When we have a **sample**, we just use some primary school mathematics to determine the sample mean:

$$\bar{x} = (1/n) \sum_{i=1}^n x_i$$

- ▶ Note that the *population* mean μ (i.e. the mean of the underlying distribution) is **not** (in general) the same as the sample mean \bar{x} . But **we expect $\mu = \bar{x}$ as $n \rightarrow \infty$** .
 - ▶ If the sample reflects the underlying distribution though (e.g. a sample of 1,2,3,4,5,6 for a fair die) then the formulas are equivalent.
-



The basics...

- ▶ The **variance** σ^2 of a distribution gives a measure of how 'spread out' the data is.
- ▶ Remember from school that we could find the variance of a **sample** by finding the differences from the mean, squaring them, summing them then dividing by the size of our sample. e.g. Imagine we throw one each of 1, 2, 3, 4, 5 and 6 on a fair dice.

X	1	2	3	4	5	6
$X-\mu$	-2.5	-1.5	-0.5	0.5	1.5	2.5
$(X-\mu)^2$	6.25	2.25	0.25	0.25	2.25	6.25

- ▶ Sum = 17.5 $\sigma^2 = 17.5 / 6 = 2.917$. Gravy...



The basics...

- ▶ The standard deviation σ is just the square root of the variance.
- ▶ Why did we use the squares of the deviations from the mean rather than say the absolutes $|x_i - \mu|$? This would lead to the **absolute deviation**. Standard deviation is generally more useful because the 'squared' is more easily differentiable than the absolute value. Additionally it can be used as a parameters in distributions like the Gaussian Distribution...
- ▶ So what about the variance for a known distribution?
- ▶ $\sigma^2 = E[(X - \mu)^2]$. We'll see a nice example of this in a sec...



The basics...

- ▶ It's imperative to distinguish between distributions based on a **sample** and the **population**.

	Sample	Population
Distribution determined by...	a finite sequence of draws from the distribution.	The underlying distribution is already known .
Mean	\bar{x}	μ

- ▶ Note that the population mean could theoretically be computed by taking an infinite number of samples, since in the limit we expect each outcome to occur proportional to its probability in the underlying distribution (e.g. for a fair dice and infinite throws, we expect the counts of each outcome to be the same).
-



► Now let's move on from the basics....



Bernoulli

- ▶ A discrete probability distribution, where we have some event with two outcomes, with a probability p of success.
- ▶ Who gives a toss? Let's say a referee. He has a biased coin.

Support →

k	$\Pr(K = k)$
1 (H/success)	p
0 (T/failure)	$1-p$

- ▶ A **Bernoulli Trial** is a single experiment in which we draw from this distribution, such that the outcome is **independent** of previous trials (i.e. The coin has no memory).
 - ▶ We use $k=1$ for success (interpret as '1 success') and $k=0$ otherwise.
-



Bernoulli

- ▶ Note that this is not the same as an arbitrary 2-event distribution, the support of the distribution is $\{0, 1\}$, not $\{T, H\}$.
- ▶ The mean is just p (using our equation for $E[X]$ we get $0 \cdot (1-p) + 1 \cdot p = p$).
- ▶ The variance is $p(1-p)$. Remember that $\sigma^2 = E[(X - \mu)^2]$?

$$\begin{aligned} E[(X - \mu)^2] &= \sum_i (x_i - \mu)^2 p(x_i) \\ &= (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p \\ &= p(1-p) \end{aligned}$$



Binomial

- ▶ But suppose more generally we have n Bernoulli Trials. The Binomial Distribution gives us a distribution over the number of 'successes'. Suppose we have 4 fair coin tosses:

k	$\Pr(K = k)$
0 (i.e. TTTT)	1/16
1	4/16
2	6/16
3	4/16
4 (i.e. HHHH)	1/16

- ▶ How do we work out $\Pr(K = k)$? Suppose again $n=4$, and that $k=2$. Let's take one possible sequence, HHTT. Clearly the probability is $p * p * (1-p) * (1-p)$ since the 4 trials were independent.
-



Binomial

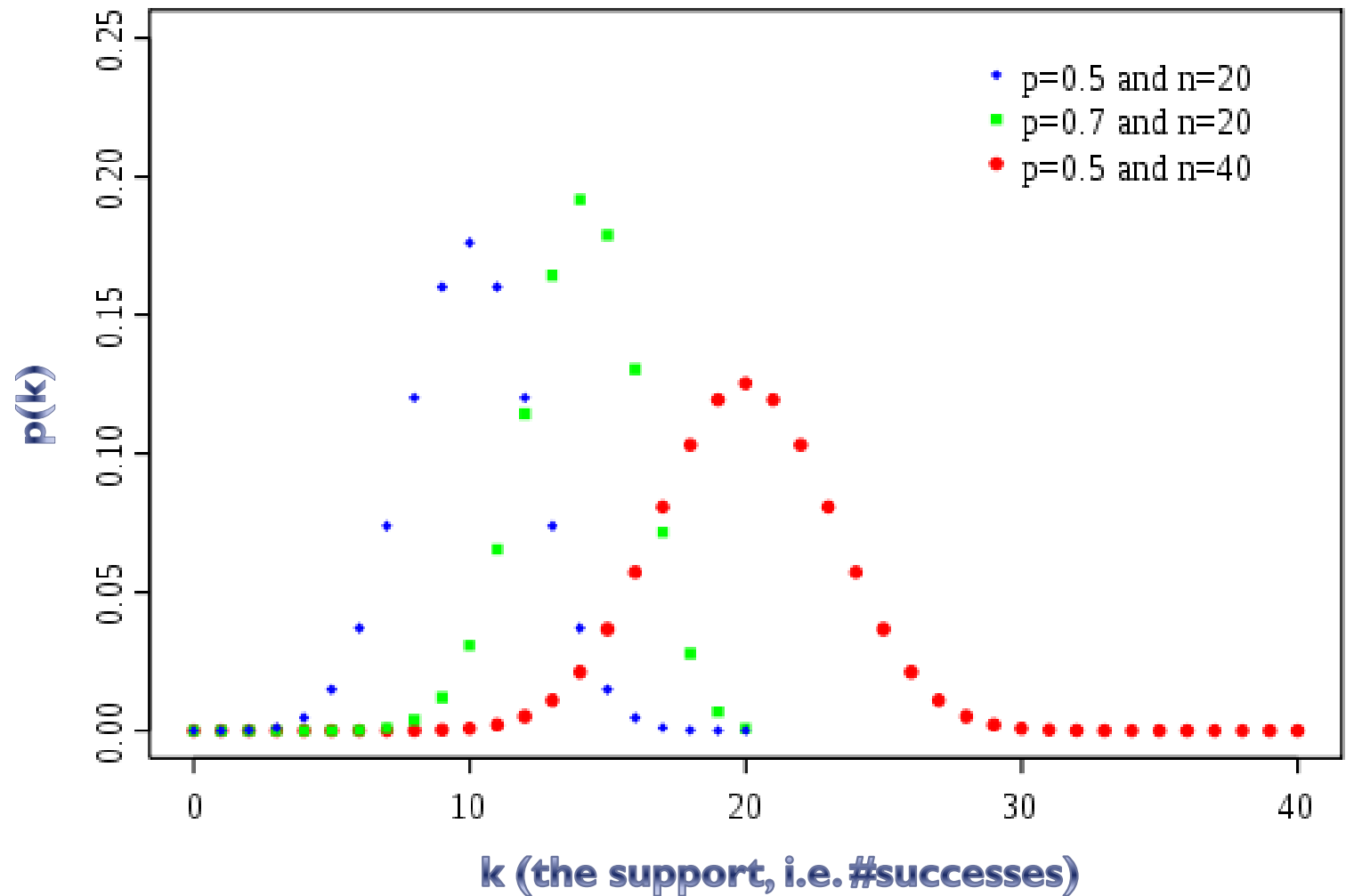
- ▶ But there's multiple ways we could have generated 2 heads and 2 tails: HHTT, HTHT, HTTH, THHT, THTH, TTTH.
- ▶ In general, for n trials and k 'successes', there are ${}^nC_k = n! / (n-k)!k!$ possible outcomes. (Intuition: If we had n different objects, there would be $n!$ ways of organising them. But they're not all different, so we need to divide by the number of ways the k successes can be arranged $= k!$, and the number of ways the failures can be arranged $= (n-k)!$)

$$\text{Therefore } \Pr(K=k) = {}^nC_k p^k (1-p)^{n-k}$$

- ▶ Unsurprisingly, the **mean is np** . If we have $n=10$ tosses of a biased coin where $p=0.1$ and do this repeatedly, it seems sensible that on average we'll see a head just the once.
- ▶ The variance is $np(1-p)$.
- ▶ Notice that when $n = 1$, we have the Bernoulli Distribution!



Binomial



Multinomial

- ▶ But suppose we wanted to **generalise** from 'success/failure' **to any number of outcomes**, e.g. a dice.



- ▶ Instead of the distribution being defined over the number of successes (e.g. #heads), we now have counts over multiple variables (i.e. for our dice, #1s, #2s, etc.). Let x_i be the count of each.
- ▶ Instead of a single parameter p , for k outcomes we have a vector (p_1, \dots, p_k) , i.e. the prob for each outcome.



Multinomial

- ▶ We pretty much have the probability for a certain set of counts (x_1, \dots, x_k) already from our Binomial distribution...
- ▶ The probability of a given sequence (e.g. 1,2,4,3,4,2) is again the product of the probabilities of each, i.e. $\prod p_i^{x_i}$ (Note: The $(1-p)$ term before was just a convenience that allowed us to specify the prob of a 2-outcome experiment as 1 parameter instead of 2.)
- ▶ As for the number of sequences that are possible with the given counts, instead of dividing $n!$ by the ways of arranging successes and failures, we generalise it to the ways of arranging each of the outcomes in the sequence.
- ▶ So (cue fanfare):

$$\begin{aligned} \Pr(X_1 = x_1 \text{ and } \dots \text{ and } X_k = x_k) \\ = [n! / \prod x_i!] \prod p_i^{x_i} \end{aligned}$$



A quick digression...

- ▶ Note the difference between the underlying parameters of the distribution and the support. The parameters are fixed whereas the support is the domain of the probability function.

f	Support	Parameters
Bernoulli	$k \in \{0, 1\}$	p
Binomial	$k \in \{0, \dots, n\}$	n, p
Multinomial	$x_i \in \{0, \dots, n\} \forall i$	$n, \langle p_1, \dots, p_n \rangle$

- ▶ Some notational notes: The support when a single parameter was k , and \mathbf{x} when a vector (i.e. a multivariate distribution).
- ▶ We often express the probability function as $f(a; b_1, b_2, \dots)$, which just means $\Pr(a \mid b_1, b_2, \dots)$, i.e. with parameters b_i .



Geometric

- ▶ The distribution is over the number of Bernoulli Trials needed to get ONE success*.
- ▶ Consider a game where with probability p we stop the game (interpret as 'succeed'), otherwise we take a random letter and add it to some word (initially blank).
- ▶ Then in this case the geometric distribution is over the length of the words we can generate.
- ▶ The probability of having k trials ending with a success is trivial:

$$\Pr(K = k) = (1-p)^{k-1} p$$

*A common variant is the number of failures before one success. The difference there is that k starts from 0 rather than 1.



Geometric

- ▶ **Where you might find it:** Probabilistic models for generating tree structures (where the geometric distribution is over the height of the tree) and words (as we just saw).
- ▶ **Mean = $1/p$.** A simple expression, but not a trivial proof:

$$E(X) = \sum_{k=1}^{\infty} (1-p)^{k-1} p \cdot k = p \sum_{k=1}^{\infty} (1-p)^{k-1} k$$

$$= p \left[\frac{d}{dp} \left(- \sum_{k=1}^{\infty} (1-p)^k \right) \right] = -p \frac{d}{dp} \frac{1}{p} = \frac{1}{p}$$

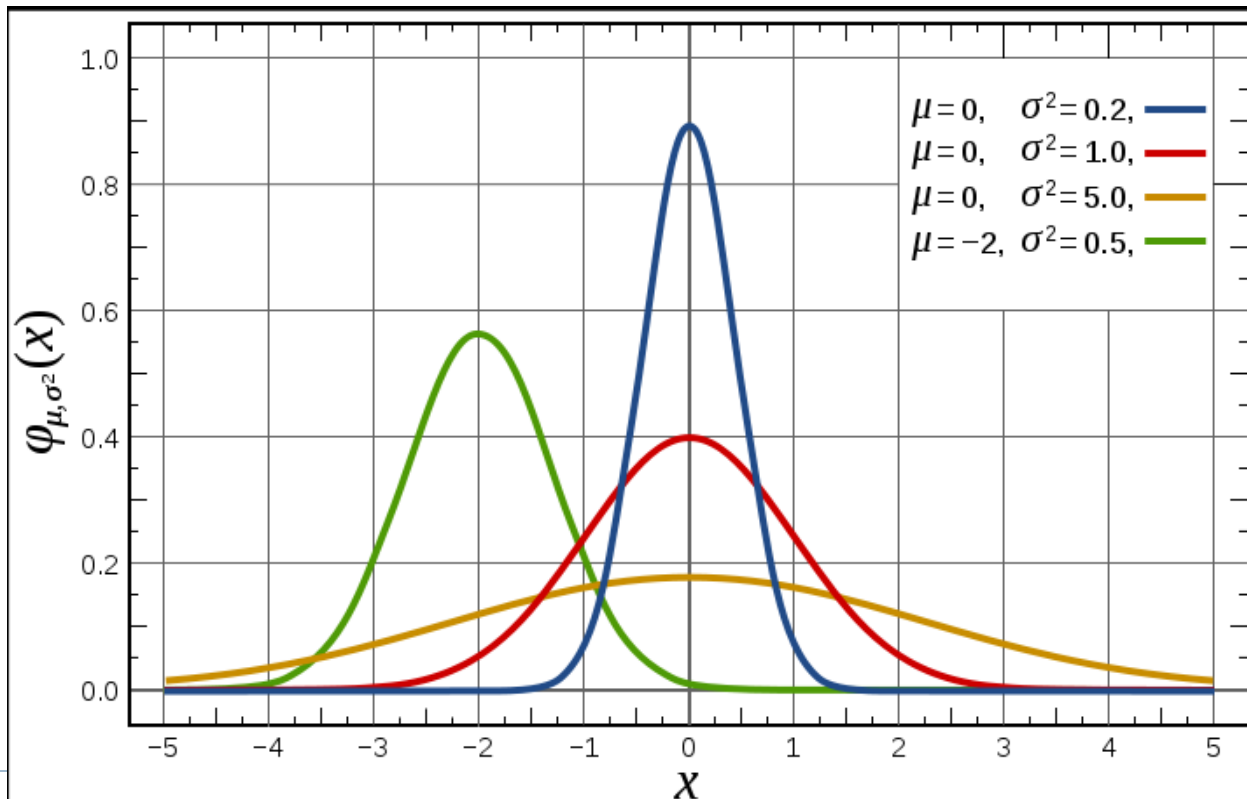
Remember that the mean $E(X) = \sum f(k) \cdot k$

- ▶ Variance = $(1-p)/p^2$
- ▶ Note that the ‘geometric’ is the same meaning as in ‘geometric series’, where you might want to find $1 + (1/2) + (1/4) + \dots$ after a certain number of additions. Each number we add becomes smaller by some factor.



Gaussian

- ▶ Describes data distributed about a mean, with a smooth ‘bell-curve’ shape.
- ▶ It is parameterised by a mean μ and a variance σ^2 .



Gaussian

- ▶ Also known as the **Normal Distribution**.
- ▶ When $\mu = 0$ and $\sigma^2 = 1$, known as **Standard N.D.**
- ▶ Unlike previous distributions which were **discrete**, this distribution is defined over **continuous** variables. Let support be $x \in (-\infty, \infty)$, i.e. a single unbounded variable.
- ▶ It is useful because the distribution has **maximal entropy**; that is, it makes as little assumption about the distribution of data as possible. Therefore useful for modelling **noise**.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



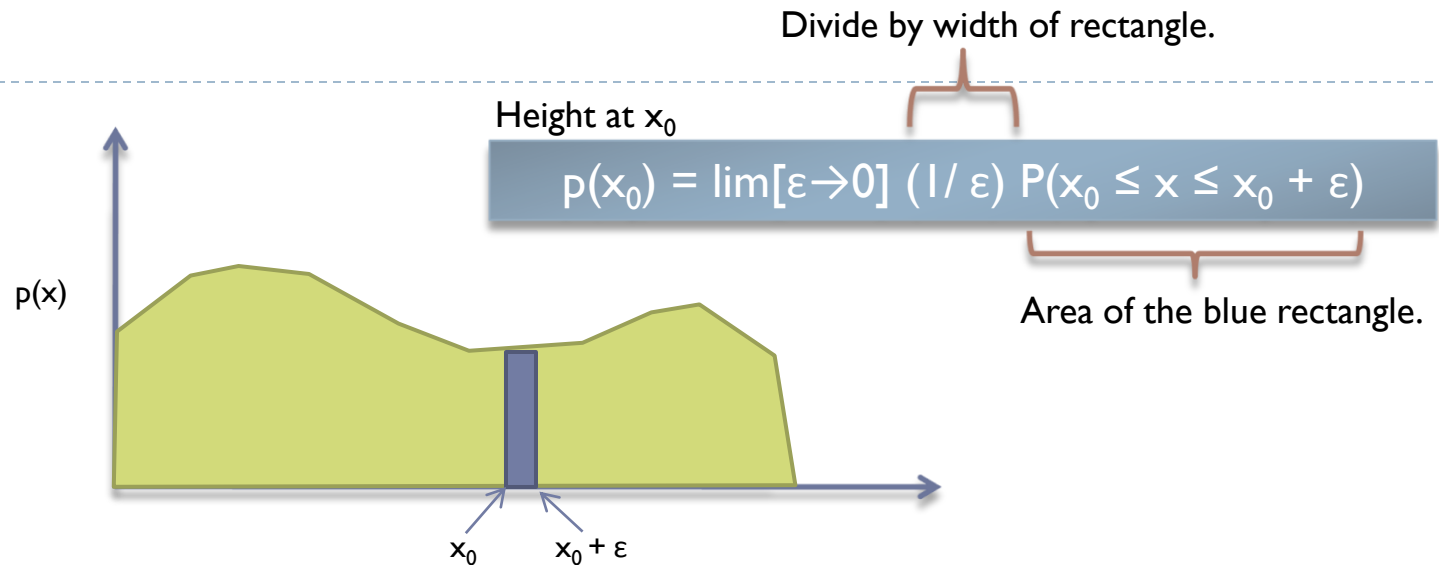
Probability Density Functions

- ▶ Previously we used **probability mass functions**, i.e. The probability at an exact point. For continuous distributions this doesn't make sense, since we have an infinitely small point.
- ▶ With a **probability density function**, consider getting the probability mass (i.e. the area) between the point x_0 under consideration and $x_0 + \varepsilon$. The width of this area is ε , so we divide by ε to get the height. We get the probability density as ε approaches 0.

$$p(x_0) = \lim[\varepsilon \rightarrow 0] (1/\varepsilon) P(x_0 \leq x \leq x_0 + \varepsilon)$$



PDFs

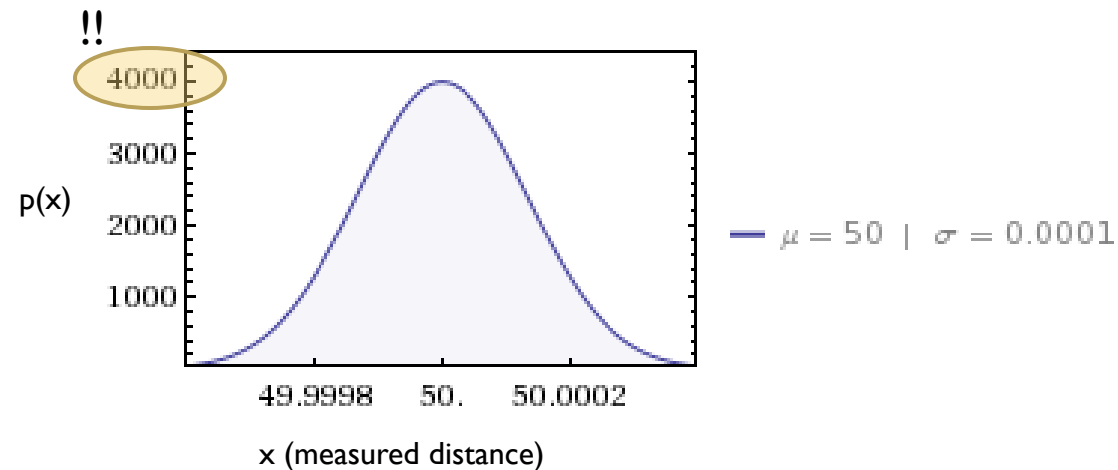


- ▶ And given we now have a continuous support, using the \sum in calculating the mean/variance no longer makes sense.
- ▶ Just replace the \sum with a \int and we're done! (recall that \int finds the area under a function, i.e. a sum of infinitely small vertical strips)

$$E[X] = \int x f(x) dx$$

PDFs

- ▶ For a probability mass function the probabilities needed to sum to 1...
- ▶ Now the integral of the probabilities (i.e. the **area under the graph**) must be 1. i.e. $\int p(x) dx = 1$
- ▶ Note also the probability density function **can give a value greater than 1.**



Imagine for example we were sensing a wall, and our reading gave a distance of 50 metres. We could use a Gaussian to represent the **actual distance**. Imagine our sensor was really really accurate: as the variance tends towards 0, we get a 'spike' of infinitely small width, but $p(x=50)$ approaches ∞ !

Gaussian

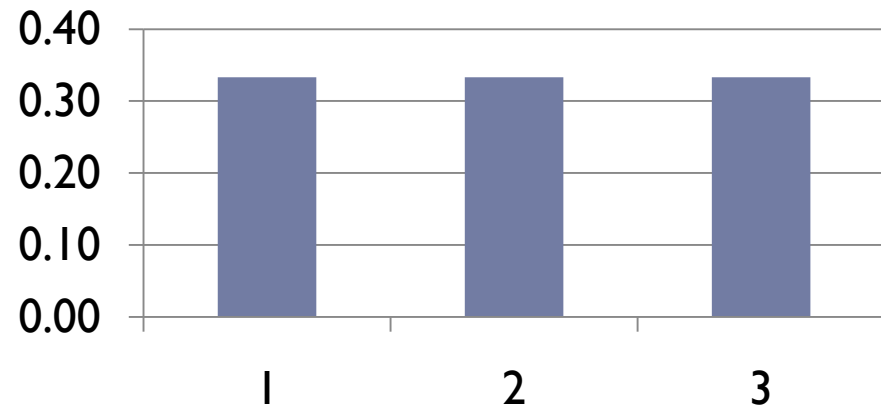
- ▶ The **Central Limit Theorem** states that the mean of a sufficiently large number of identically distributed random variables, each with mean μ and variance σ^2 , will be approximately normally distributed.
- ▶ This is rather awesome, because **even if we don't know the underlying distribution, we know the distribution of the mean.**



Gaussian

- ▶ The **sum** of identical random variables also gives a Gaussian Distribution (the mean for a sample just involved taking the sum, plus a division by a constant). Let's see an example!
- ▶ We have a simple uniform distribution for X like so:

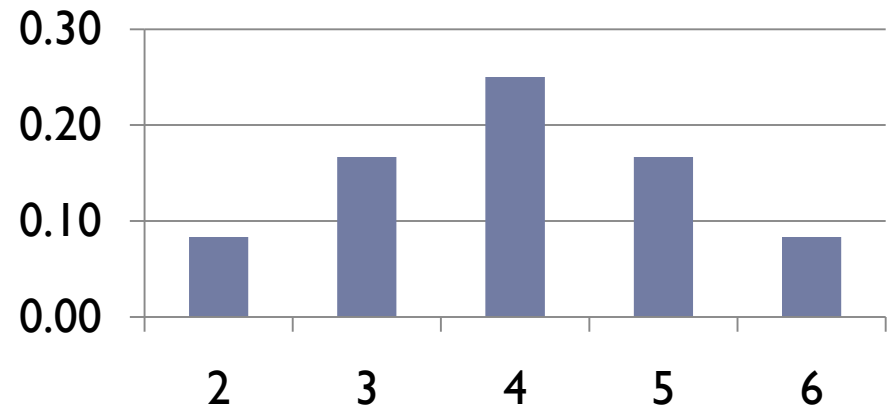
X	$\Pr(X = x)$
1	1/3
2	1/3
3	1/3



Gaussian

- ▶ Now let's take the probability mass function for the sum of 2 terms:

X	Pr(X = x)
2 (1+1)	1/9
3 (1+2, 2+1)	2/9
4 (1+3, 2+2, 3+1)	3/9
5 (2+3, 3+2)	2/9
6 (3+3)	1/9

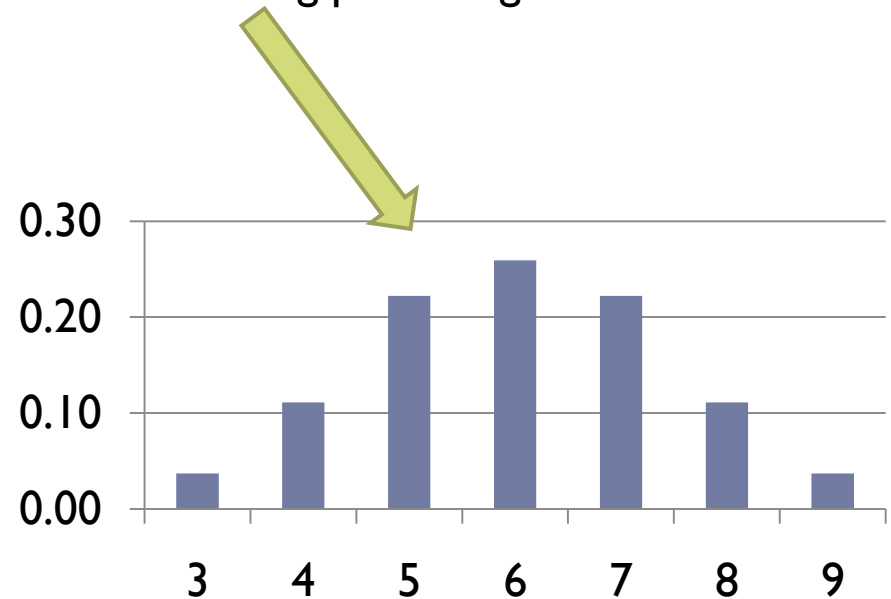


Gaussian

► 3 terms:

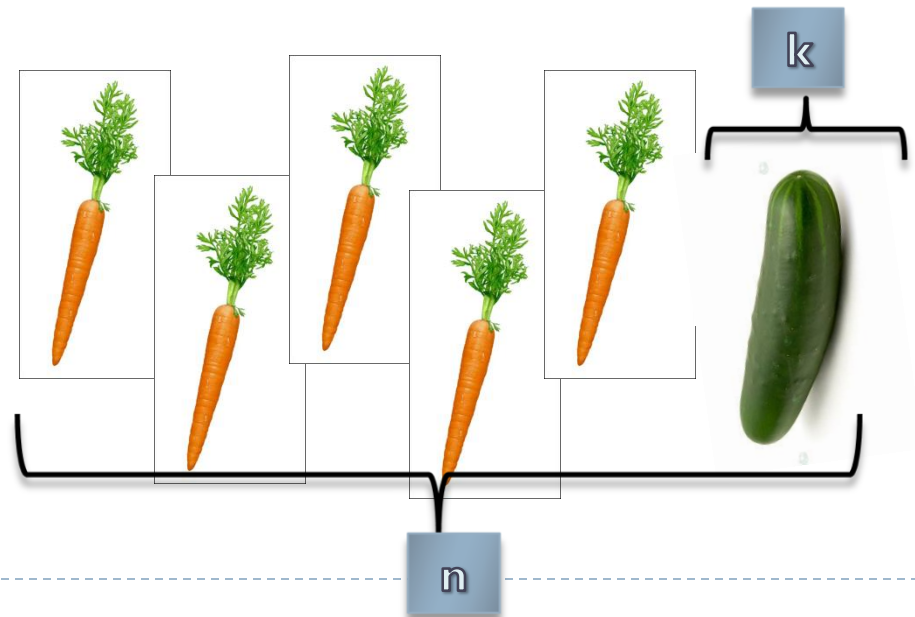
X	$\Pr(\mathbf{X} = \mathbf{x})$
3	$1/27$
4	$3/27$
5	$6/27$
6	$7/27$
7	$6/27$
8	$3/27$
9	$1/27$

This is looking promising!



Gaussian

- ▶ Another example, let's hop to it...
- ▶ Suppose we have a rabbit that picks up what he thinks are n carrots. However, k of these are not actually carrots, but cucumbers.
- ▶ We can think of this as n separate Bernoulli Trials, where 'success' in this case actually means our rab has made a grave carrot-related mishap.



Gaussian

- ▶ For these n **samples** of our Bernoulli Distribution, we can easily estimate the carrot misclassification probability (or c.m.p. for short) as k/n (using our sample mean eq).
- ▶ We know the mean of a Bernoulli Distribution is p , so our mean is the same as $p = k/n$.
- ▶ But we can't be sure about our choice of p^* . If the rabbit chose 24 objects and 12 were carrots, it might actually be the case that $p = 0.6$ for example.
- ▶ If we were to randomly sample from our Bernoulli Distribution a number of times, how can we describe the uncertainty over the mean, i.e. p ?

* p is known as an **estimator**, as it used to estimate (via use of a sample) an unknown **population parameter**, i.e. a parameter for the underlying distribution.



Gaussian

- ▶ Because of the Central Limit Theorem, since we have identically distributed random variables (each random variable was the same Bernoulli Trial), **the distribution over the mean/ p can be described using the Gaussian Distribution as $n \rightarrow \infty$** (i.e. the case where the rabbit gets incredibly fat).
 - ▶ If we had a **finite** number of trials N , where $N > n$, we could use the Binomial Distribution to define the uncertainty over p . We could initially plot it for k so we have the probability for each number of misclassifications, then replace each k in the support vector with k/N so our uncertainty is over p instead.
-



Gaussian

- ▶ Therefore the Gaussian Distribution is simply the Binomial Distribution when $n \rightarrow \infty$. This makes me a happy bunny...
- ▶ It's *bunnies all the way down...** But let's not rabbit on...

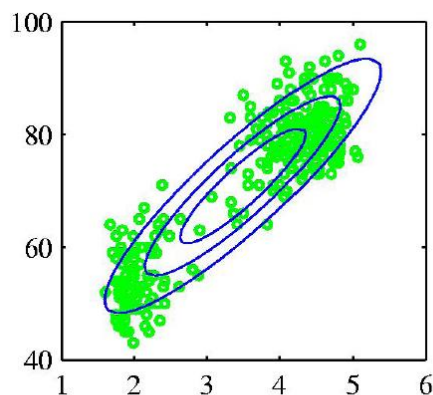


* Note: CLG in-joke.

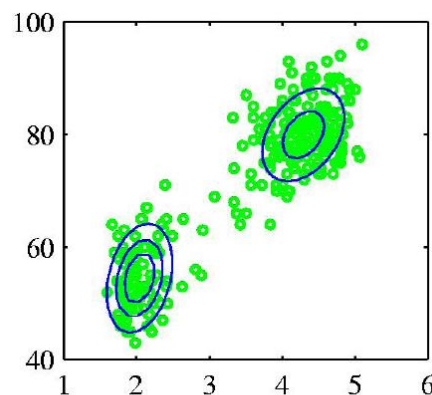


Gaussian

- ▶ One more quick common example usage... **Gaussian Mixture Models** allow us to cluster data, but instead of a point being a member of one cluster or another ('hard' assignment), we can associate a probability with the point being a member of each cluster (i.e. 'soft' assignment).



Single Gaussian



Mixture of two Gaussians

- ▶ How do we define a Gaussian in more than 1 dimension?

Gaussian

- ▶ **Covariance** gives a measure of how two random variables change together. For 2 variables X and Y :

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

- ▶ Notice that variance is just a special case of this, when the 2 variables are the same.
- ▶ If two variables are **independent**, their covariance is 0.
- ▶ If we have D random variables X_1, \dots, X_D , (for example, x and y in 2D space) then a **covariance matrix** Σ is a symmetric matrix where each element (i, j) is just $\text{Cov}(X_i, X_j)$.



Gaussian

$$\Sigma = \begin{bmatrix} \mathbb{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & \mathbb{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\ \mathbb{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \mathbb{E}[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & \mathbb{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \mathbb{E}[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & \mathbb{E}[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}.$$

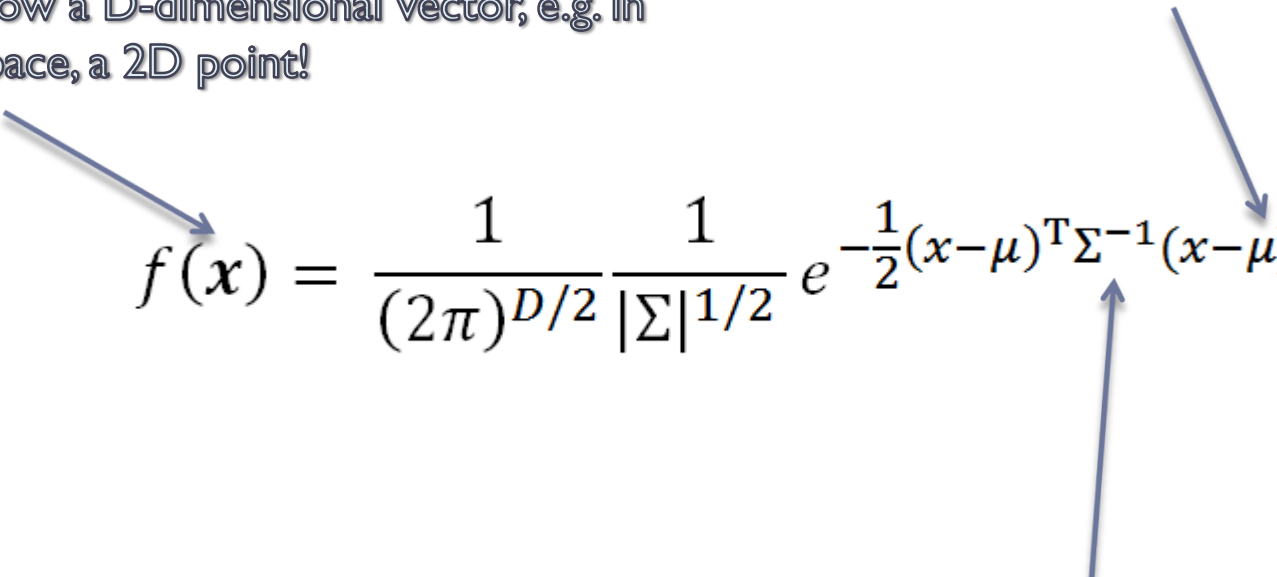


Gaussian

- ▶ Then a multivariate Gaussian is:

x is now a D -dimensional vector, e.g. in 2D space, a 2D point!

The mean μ is also now D -dimensional



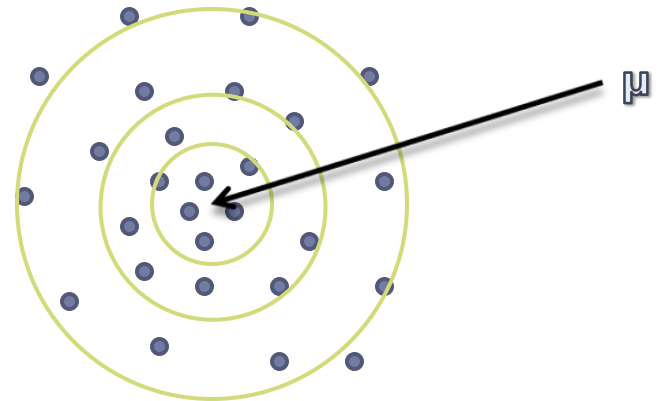
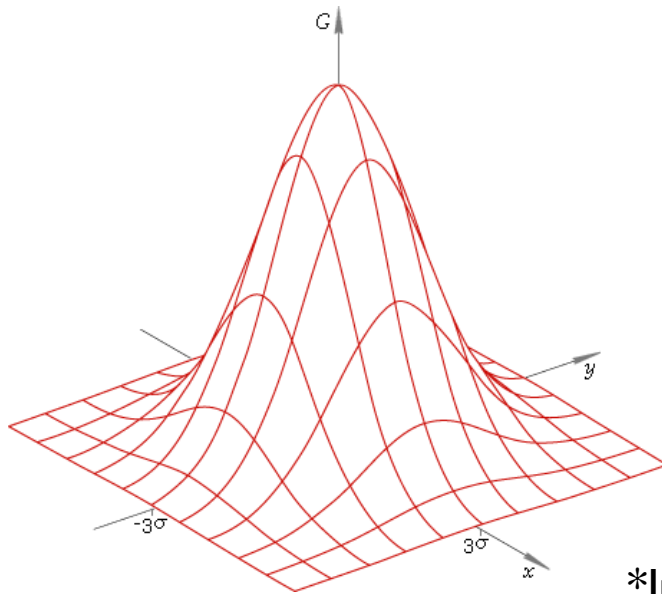
The diagram consists of three blue arrows pointing towards the formula. One arrow originates from the text 'x is now a D-dimensional vector, e.g. in 2D space, a 2D point!' and points to the variable x in the exponent. A second arrow originates from the text 'The mean μ is also now D-dimensional' and points to the variable μ in the exponent. A third arrow originates from the text 'The inverse of the covariance matrix is often called the concentration matrix or the precision matrix.' and points to the Σ^{-1} term in the exponent.

$$f(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

The inverse of the covariance matrix is often called the concentration matrix or the precision matrix.

Gaussian

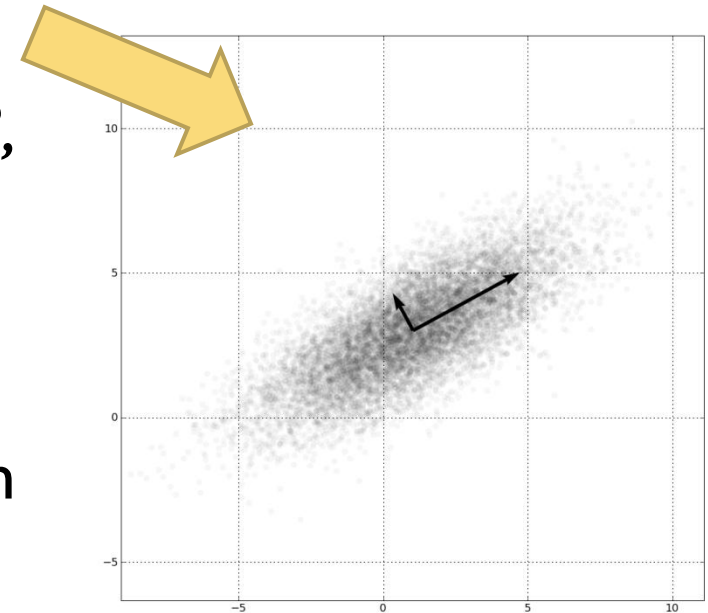
- ▶ If we use the identity matrix I for the covariance matrix (i.e. The variables don't 'change' with any other variable), then if we plot a contour* map over say 2D space for the distribution, we get perfect circles.



*Just like on geographical maps, a contour is a line where the value of the function is invariant.

Gaussian

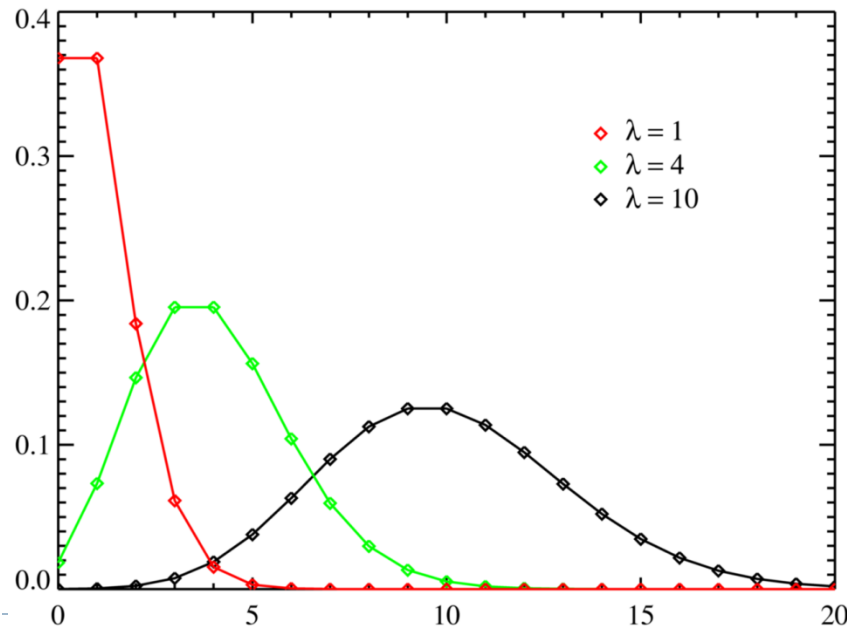
- ▶ A multivariate Gaussian distribution centred at $(1,3)$ with a standard deviation of 3 in roughly the $(0.878, 0.478)$ direction and of 1 in the orthogonal direction.
- ▶ The directions of the arrows show the 'axis' of the distribution. We can determine these by finding the **eigenvectors** of this covariance matrix (feel free to ask me what these are!).



Because the x and y components covary, the x and y components of variance do not fully describe the distribution. A 2×2 covariance matrix is needed.

Poisson

- ▶ *Oh my cod*, not another discrete non-finite distribution!
- ▶ Expresses the probability of a given number of events happening in a fixed amount of time, when there is a known **average rate λ** at which the events occur.



(Note: This is a discrete distribution, so the connecting lines are merely a guide.)

Poisson

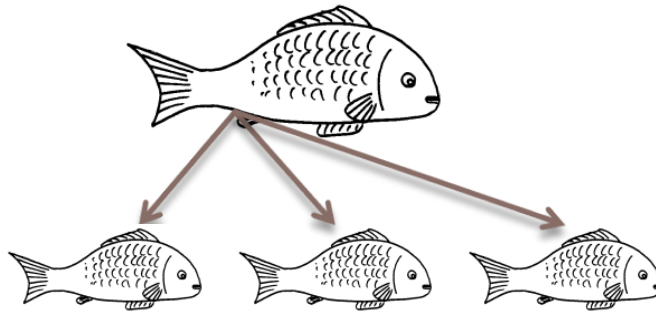
- ▶ Examples: num times a web server is accessed per minute, number of slides I get through in one hour...
- ▶ Surprisingly a specific case of the Binomial Distribution. Recall that for a BD, np gives us the mean number of successes.
- ▶ If $n \rightarrow \infty$ and p is small, then we have a Poisson Distribution if we set the rate $\lambda = np$. Known as the **law of rare events**. But the proof has no 'place' here.

$$f(k; \lambda) = (\lambda^k * e^{-\lambda}) / k!$$

What is the prob of k events occurring within our time period, given an average rate of λ

Poisson

- ▶ **Sometimes it is appropriate to choose a distribution simply because of its shape**, rather than conceptually what it means... (it's not as fishy as it sounds)
- ▶ The Poisson Distribution is useful if we wish to randomly generate a value that has some mean > 0 , must be > 0 , and the probability tails towards 0 as the value drifts above the mean. Example: generating a random number of children for a tree.

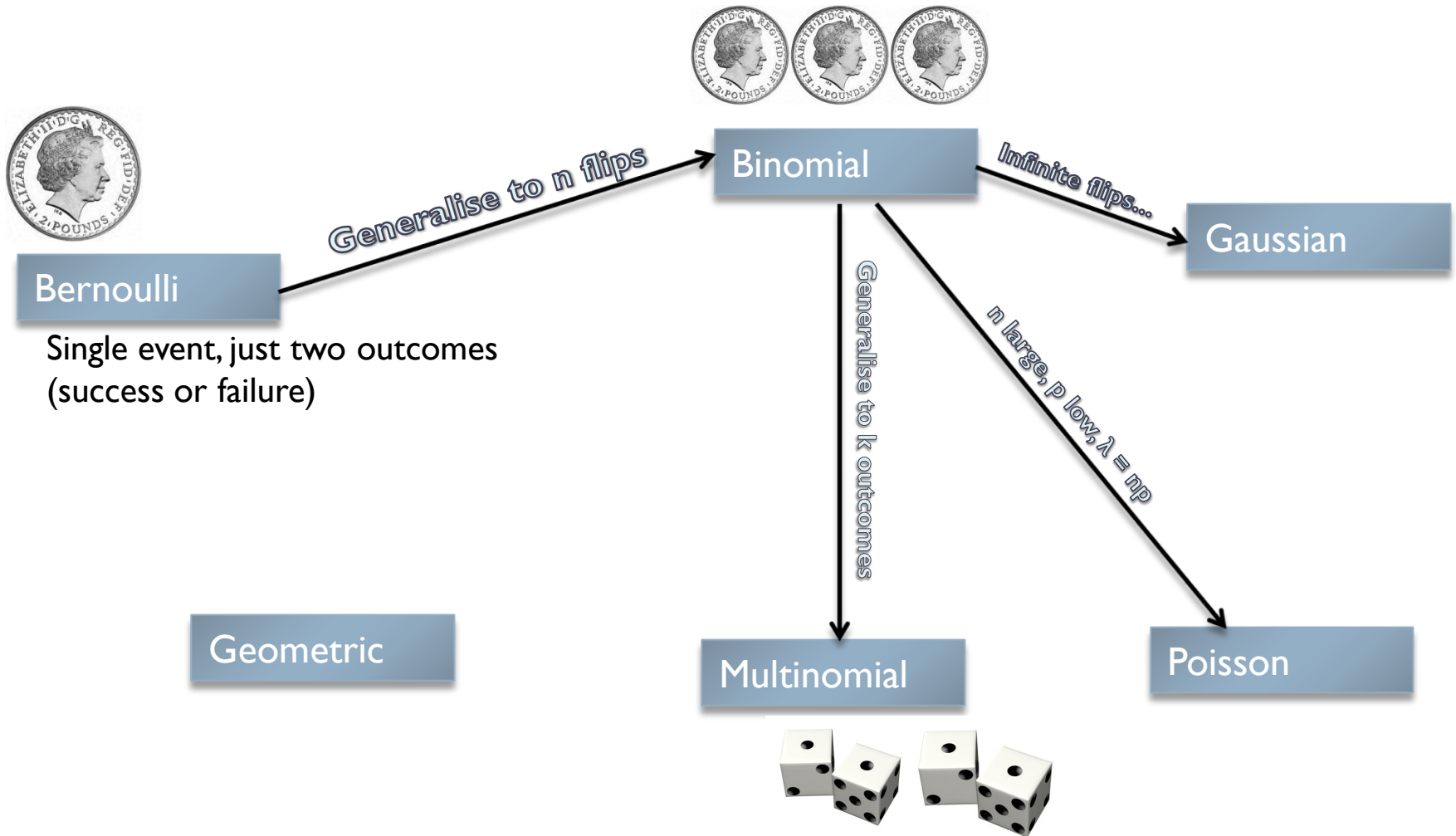


Summary

Distribution	Support	Parameters	p.m.f. / p.d.f.
Bernoulli	$k \in \{0, 1\}$	p	$p^k(1-p)^{1-k}$
Binomial	$k \in \{0, \dots, n\}$	n, p	${}^nC_k p^k (1-p)^{n-k}$
Multinomial	$x_i \in \{0, \dots, n\} \forall i$	$n, \langle p_1, \dots, p_n \rangle$	$[n! / \prod x_i!] \prod p_i^{x_i}$
Gaussian $\mathcal{N}(\mu, \sigma^2)$	$x_i \in [-\infty, \infty]$	μ, σ^2	$f(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{ \Sigma ^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$
Geometric	$k \in \{1, 2, \dots\}$	p	$(1-p)^{k-1} p$
Poisson $Pois(\lambda)$	$k \in \{0, 1, 2, \dots\}$	λ	$(\lambda^k * e^{-\lambda}) / k!$

This is generally the notation we use for the distribution – particularly when we are drawing from it, e.g. $X \sim Pois(\lambda)$ or $X \sim N(\mu, \sigma^2)$.

Summary





Fini

End of Part I