PART 1: AN INTRODUCTION TO STATISTICS AND INFERENCE

Matthew Pitkin

GraWIToN School

25 March 2015

University of Glasgow

OVERVIEW

Part 1: *An introduction to statistics and inference*

- Rules of probability
- Bayes' theorem
- Important probability density functions (pdf)
- Moments of pdfs

Part 2: *Frequentist statistical building blocks*

- Statistics and estimators (parameter estimation)
- Least squares fitting
- Maximum likelihood
- Hypothesis tests
- Goodness of fit

Part 3: *An introduction to Bayesian inference*

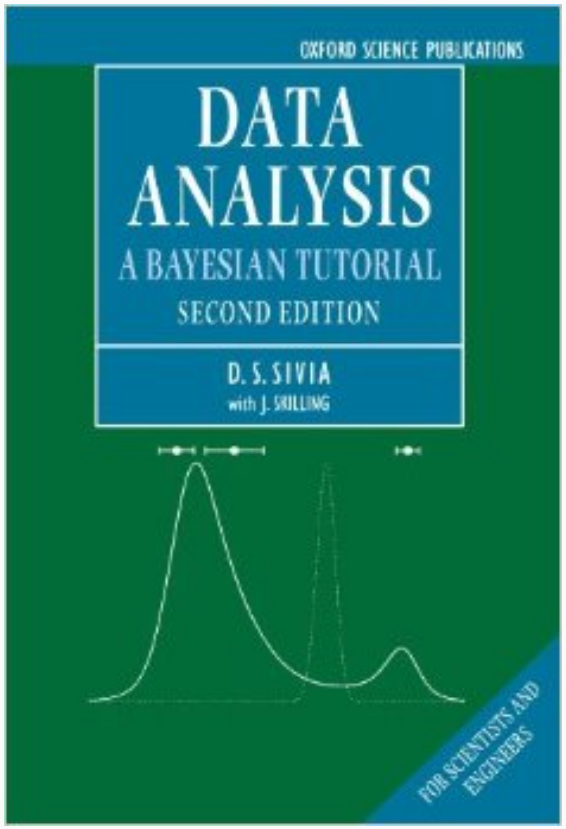
- Bayesian parameter estimation
- Assigning probabilities
- Bayesian hypothesis testing
- The Gaussian approximation

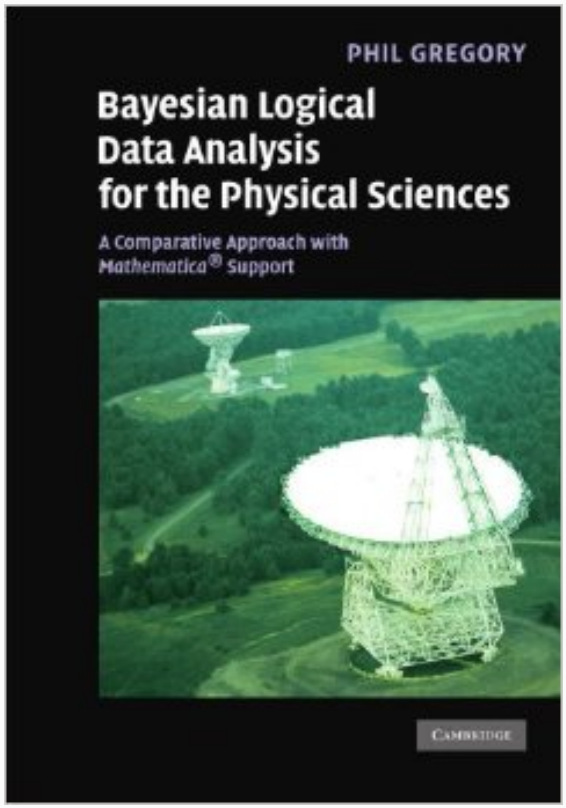
Part 4: *Practical Bayesian methods*

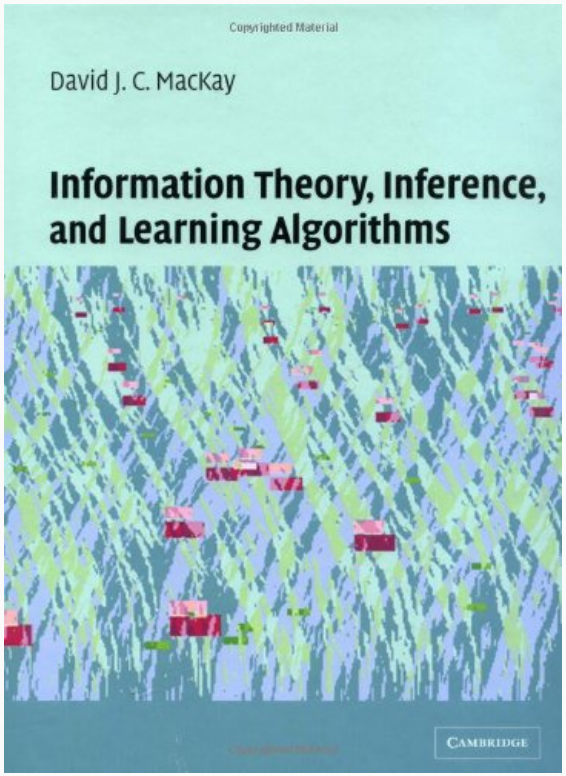
- An introduction to MCMC

This lecture course is based heavily on the SUPA Advanced Data Analysis course by Prof. Martin Hendry.

There are many textbooks on statistics (and Bayesian statistics in particular), but three I can recommend are:







- Bernoulli (1713) wondered how the mechanics of deductive logic applied to games of chance could be applied to inductive logic problems
- Bayes (1763 - postumous) provided an answer (Bayes' formula) to Bernoulli's questions
- Laplace (1812) independently derived Bayes' theorem in the more common form we know today

Probabilities represent a *plausibility* or *degree-of-beilef* of a proposition given the evidence at hand.

“Probability theory is nothing but common sense reduced to calculation” - Laplace

DEDUCTIVE AN INDUCTIVE REASONING

- Deductive logic
  - A cause (e.g. a theory) can lead to many effects or outcomes (predictions of the theory)
- Inductive logic
  - Observations can be used to *infer* the *plausibility* their of possible causes (e.g. different theories or models)
  - Also known as *inverse probability* - learning about a model/theory from the data/observations

DEDUCTIVE LOGIC: AN EXAMPLE

We have a set of statements:

A. All observed gravitational waves in the southern hemisphere are circularly polarised
B. A gravitational wave is observed to originate in the southern hemisphere
C. A gravitational wave is observed to be circularly polarised

Suppose statement A (our theory) *is* true:

- if statement B is true, then statement C is also true
- if statement C is false, then statement B is also false

Statement C is a logical consequence of A and B.

INDUCTIVE LOGIC: AN EXAMPLE

A. All observed gravitational waves from the southern hemisphere are circularly polarised
B. A gravitational wave is observed to originate in the southern hemisphere
C. A gravitational wave is observed to be circularly polarised

What can we say about B *if* A and C are true?

INDUCTIVE LOGIC: AN EXAMPLE

A. All observed gravitational waves from the southern hemisphere are circularly polarised
B. A gravitational wave is observed to originate in the southern hemisphere
C. A gravitational wave is observed to be circularly polarised

What can we say about B *if* A and C are true?

Note that Statement A doesn't say that all circularly polarised gravitational waves originate in the southern hemisphere, but we might say that

- if C is true then B is more plausible

PLAUSIBLE REASONING

In the 1940s and 50s Cox, Polya and Jaynes formalised the mathematics of inductive logic as plausible reasoning

Probability measures our degree of belief that something is true

Axioms of Cox [1] [2] combined with Boolean algebra, are sufficient to uniquely specify the theory of probability:

1. Degree of belief must be a real number between 0 and 1
2. "*The probability of an inference on given evidence determines the probability of its contradictory on the same evidence.*"
3. "*The probability on given evidence that both of two inferences are true is determined by their separate probabilities, one on the given evidence, the other on this evidence with the additional assumption that the first inference is true.*"

1. Degree of plausibility must be a real number between 0 and 1
   - $P(X) = 1 \rightarrow$ we are *certain* that X is true
   - $P(X) = 0 \rightarrow$ we are *certain* that X is false
2. Belief in a proposition and its negation are related
   - $P(\text{not } A) = f(P(A))$
3. Belied in a pair of propositions x, y (x *and* y) is related to the belief in the conditional proposition $x|y$ ("x given y is true")
   - $P(x, y) = f(P(x|y), P(y))$

The degree of belief will always depend on available background information or assumptions, I.

E.g. we write $P(X|I)$ for the probability that X is true given I, where the | denotes conditional probability:

- our state of knowledge about X is *conditioned* by the background information/assumptions I

Throughout this course we will use the notation that 'P' represents a probability, whilst 'p' represents a probability density function, such that

$$P(x_1 < X < x_2|I) = \int_{x_1}^{x_2} p(x|I)\mathrm{d}x.$$

The rules for probabilities of propositions are inherited from classical logic and Boolean algebra:

- *Law of Excluded Middle* $P(A \text{ or not}(A)) = 1$
- *Law of Non-contradition* $P(A \text{ and not}(A)) = 0$
  - i.e. $P(A) + P(\text{not } A) = 1$ (the **sum rule**)
- *Association*
  - $P(A, [B, C]) = P([A, B], C)$
  - $P(A \text{ or } [B \text{ or } C]) = P([A \text{ or } B] \text{ or } C)$
- *Distribution*
  - $P(A, [B \text{ or } C]) = P(A, B \text{ or } A, C)$
  - $P(A \text{ or } [B, C]) = P([A \text{ or } B], [A \text{ or } C])$

RULES OF PROBABILITY

- *Commutation*
  - $P(A, B) = P(B, A)$
  - $P(A \text{ or } B) = P(B \text{ or } A)$
- *Duality* (De Morgan's Theorem)
  - $P(\text{not } [A, B]) = P(\text{not}(A) \text{ or not}(B))$
  - $P(\text{not } [A \text{ or } B]) = P(\text{not}(A), \text{not}(B))$

Note that you may see other notation for probabilities expressed with Boolean logic (this list is not exhaustive)

- Negation (A is false)
  - $P(\text{not } A)$, or $P(\bar{A})$, or $P(\neg A)$
- Logical product (both A *and* B are true)
  - $P(A, B)$, or $P(AB)$, or $P(A \text{ and } B)$, or $P(A \wedge B)$
- Logical sum (at least one of A *or* B is true)
  - $P(A + B)$, or $P(A \text{ or } B)$, or $P(A \vee B)$

From these axioms we can derive:

- The **(Extended) Sum Rule**
  - $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
- The **Product Rule**
  - $P(A \text{ and } B) = P(A)P(B|A) = P(B)P(A|B)$

These rules apply to probabilities P and also probability density functions (pdfs) p.

Simple demonstration of the extended sum rule.

What is the probability that a card drawn from a standard deck of cards is a spade *or* an ace?

We have $P(\spadesuit) = 13/52 = 1/4$ and $P(\text{ace}) = 4/52 = 1/13$, and $P(\spadesuit \text{ and ace}) = 1/(4 \times 13) = 1/52$. It is reasonably obvious that for $P(\spadesuit \text{ or ace})$ we want to sum the probabilities for both cases, however they both contain the case where $P(\spadesuit \text{ and ace})$, so we have to remove one of those instances

$$P(\spadesuit \text{ or ace}) = \frac{13 + 4 - 1}{52} = \frac{16}{52}$$

From the product rule comes

$$\boxed{P(B|A,I) = \frac{P(A|B,I)P(B|I)}{P(A|I)}}$$

This is Bayes' theorem.

From now on we will stick to using $P(A,B|I)$ to denoted the probability of A *and* B, where we have also explicitly added the conditioning on background information I.

BAYES' THEOREM

Bayes theorem can be cast in terms of a **model** and some observations, or **data**. It tells us how to update our degree of belief about our model based on new data.

$$\underbrace{P(\text{model}|\text{data}, I)}_{\text{Posterior}} = \frac{\overbrace{P(\text{data}|\text{model}, I)}^{\text{Likelihood}}\overbrace{p(\text{model}|I)}^{\text{Prior}}}{\underbrace{P(\text{data}|I)}_{\text{Evidence}}}$$

We can calculate the these terms (e.g. analytically or numerically on a computer).

BAYES' THEOREM

- **Prior**: what we knew, or our degree of belief, about our model before taking data
- **Likelihood**: the influence of the data in updating our degree of belief
- **Evidence**: the "*evidence*" for the data, or the likelihood for the data *marginalised* over the model (we'll explore this later, but at the moment note it as a constant normalisation factor for the posterior)
- **Posterior**: our new degree of belief about our model in light of the data

A degree of belief was the original way that Bayes and Laplace interpreted probabilities.

But, throughout much of the 20th century the frequentist approach has dominated. Reasons included:

- many saw the idea of probability as a *degree of belief* too subjective
- too slow/difficult to calculate

whereas the frequentist approach was

- *supposedly* objective
- provided simple "cookbook" of procedures to follow

FREQUENTIST APPROACH

Rather than being a *degree of belief* (or plausibility) this approach states

- Probability = ‘long run relative frequency’ of an event

which, *in principle*, could be measured objectively.

If rolling a die what is the probability of rolling a one?

If the die is 'fair' we expect $P(1) = P(2) = \ldots = p(6) = 1/6$.

However, imagine we don't know the die is fair. How would we go about finding these probabilities and seeing if it is?

We can imagine rolling the dice M times, and count the number of each outcome, so we define

$$P(1) = \lim_{M \to \infty} \frac{n(1)}{M}, \text{ and if } P(1) = 1/6 \text{ the die is 'fair'}$$

What do we do to estimate probabilities in the frequentist approach if $M \neq \infty$? I.e. what do we do if we just have one set of observations?

We will discuss some of the *machinary* to do this later.

FREQUENTIST VS. BAYESIAN APPROACH

- Frequentist
  - data are a **repeatable** random sample
  - true parameter **remain fixed** during this repeatable process
- Bayesian
  - data are an observation from the realised sample
  - parameters are unknown and described probabilistically

The supposed objectivity of the frequentist approach is illusory. All statistical models, be they Bayesian or Frequentist, contain assumptions, but it is important to be up front about them.

$$\text{Subjective} \neq \text{arbitrary}$$

If we have a set of M discrete propositions $\{x_k : k = 1, \ldots, M\}$, conditional on another proposition y, that cover all possibilities, then

$$\sum_{k=1}^{M} P(x_k|y, I) = 1$$

Applying the product rule: $P(x_k, y|I) = P(x_k|y, I)P(y|I)$, we get

$$\sum_{k=1}^{M} P(x_k, y|I) = \underbrace{\left[\sum_{k=1}^{M} P(x_k|y, I)\right]}_{=1} P(y|I)$$

So, the probability of y summed over xs, i.e. the marginalised probability for y, is

$$P(y|I) = \sum_{k=1}^{M} P(x_k, y|I)$$

This extends to the continuum limit in x

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

where $p(y|I)$ and $p(x, y|I)$ are *probability density functions* (pdfs).

To calculate a probability from a pdf we can find the area under it

$$P(a \leq x \leq b|I) = \int_a^b p(x|I)\mathrm{d}x$$

where we also have the normalisation condition that

$$\int_{-\infty}^{\infty} p(x|I)\mathrm{d}x = 1.$$

In the Bayesian approach the probabilities defined above can be for any proposition, however in the frequentist approach probabilities only apply to discussion of **random variables**.

- **Random variable** (RV) is "a quantity that can meaningfully vary throughout a series of repeated experiments"
  - E.g. a measured physical quantity (such as height, weight, gravitational wave amplitude) which contains random errors

The *random variable* therefore could represent data (and as we will discuss in Parts 2 & 3 could be conditional on other parameters).

Discrete pdfs involve processes where the *random variable* can only be a positive integer.

Poisson pdf: for data involving number counts

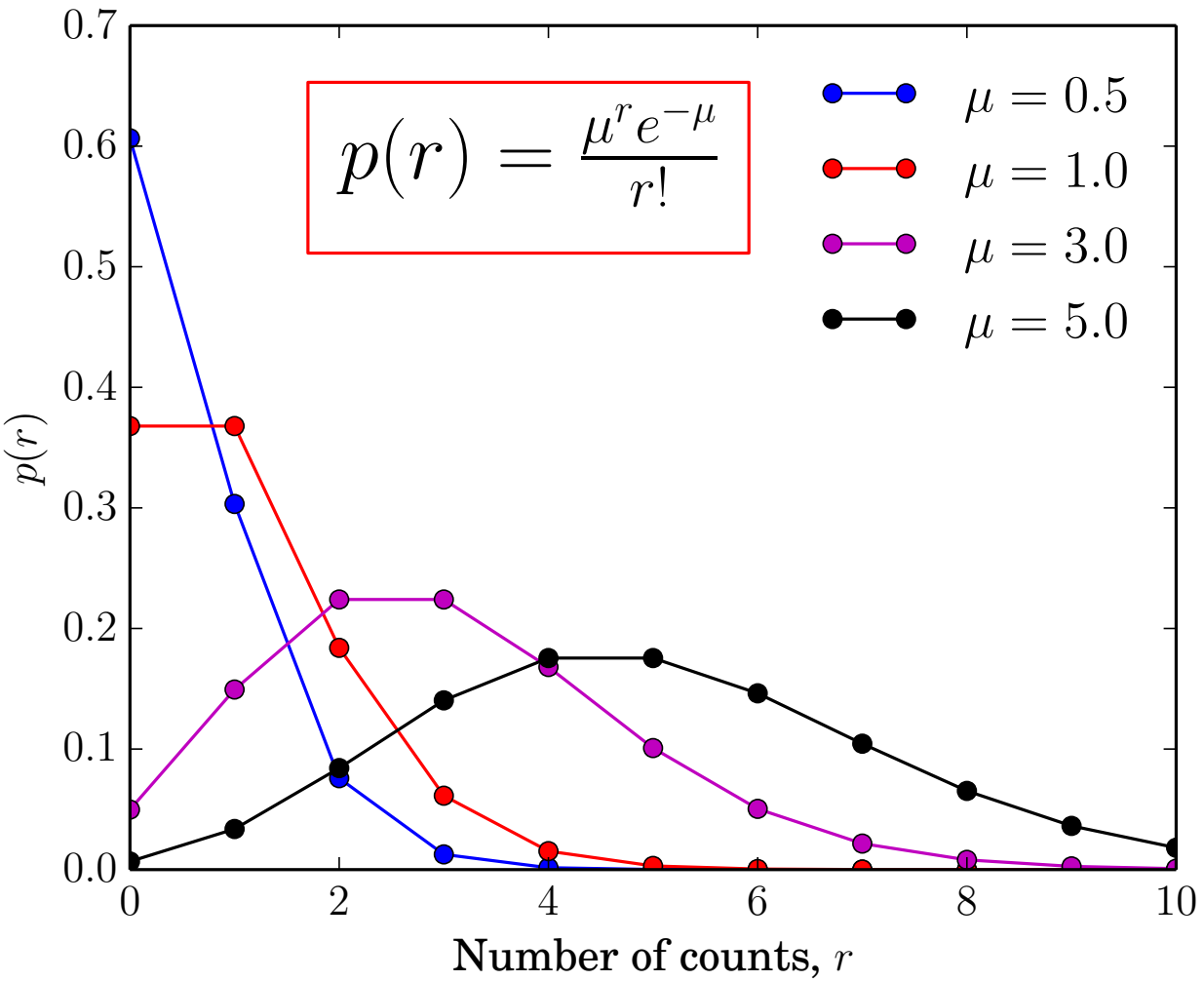
- e.g. number of photons per second counted by a CCD

Probability of a certain number of detections, r:

$$\boxed{p(r) = \frac{\mu^r e^{-\mu}}{r!}}$$

Poisson pdf assumes detections and independent and the rate μ is constant.

SOME IMPORTANT PDFS: DISCRETE PDF



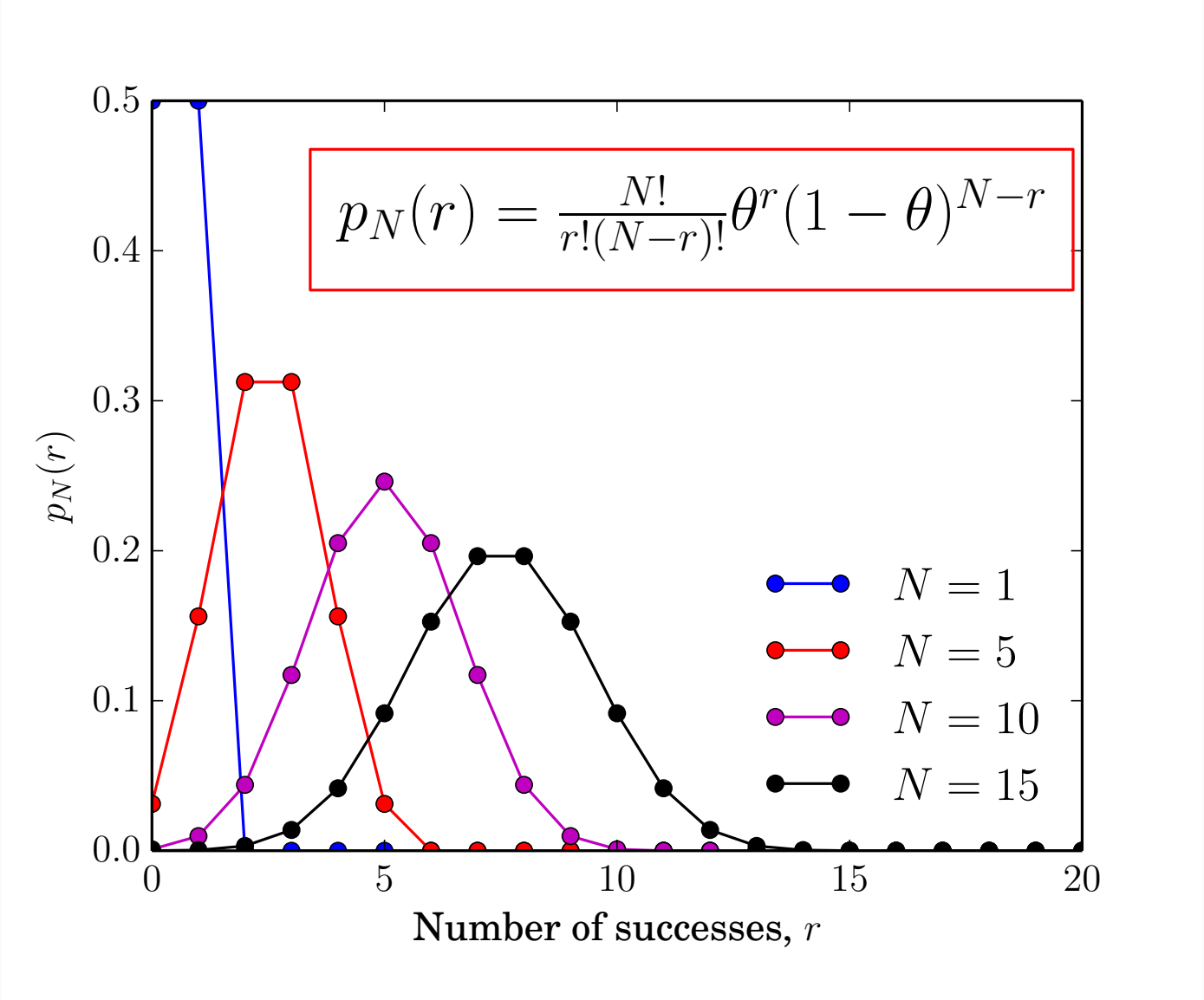
Binomial pdf: number of successes from N observations for two mutually exclusive outcomes (i.e. 'heads' or 'tails' from a coin toss)

Probability of a certain number of successes, r:

$$\boxed{p_N(r) = \frac{N!}{r!(N-r)!}\theta^r(1-\theta)^{N-r}}$$

where θ is the probability of 'success' for a single outcome, e.g. the coin comes up heads (which for a fair coin means $\theta = 0.5$).

SOME IMPORTANT PDFS: BINOMIAL PDF



In the discrete cases

$$\sum_{r=0}^{\infty} p(r) = 1,$$

so p is the probability of a given outcome.

E.g. for the binomial case, given $\theta = 0.5$ as the chance of getting a head in a single coin toss is, the probability $p_{N=1}(r = 1) = 0.5$.

Continuous pdfs are used when a *random variable* can be defined over a continuum of possible real values.

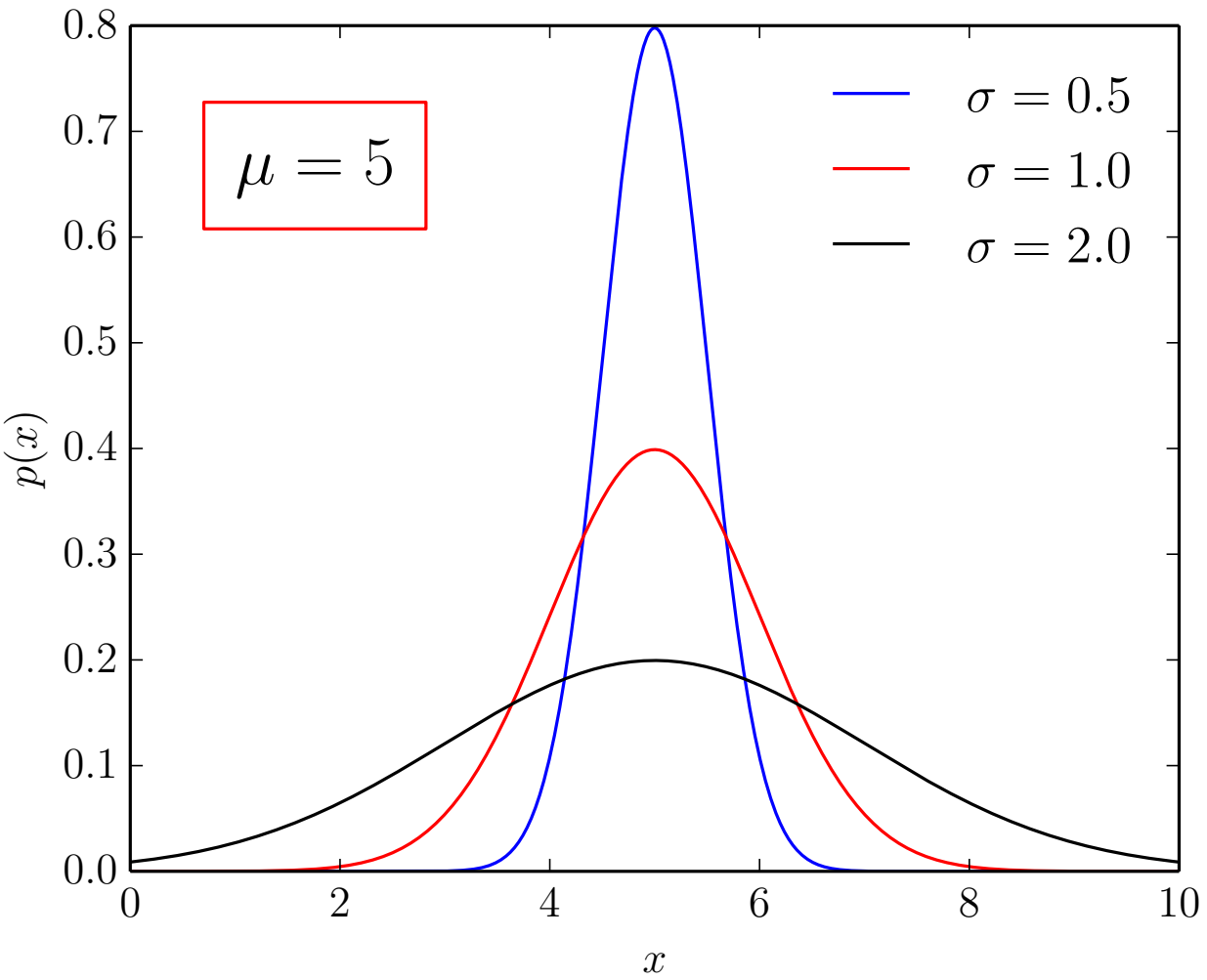
Uniform pdf: probability is uniform within a range and zero elsewhere

$$p(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b \\ 0, & \text{otherwise} \end{cases}$$

Central, Normal, or Gaussian pdf: probability has a bell-shaped distribution about a mean value μ, with a width σ (the standard deviation)

$$\boxed{p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \equiv N(\mu, \sigma^2)}$$

SOME IMPORTANT PDFS: CONTINUOUS CASE



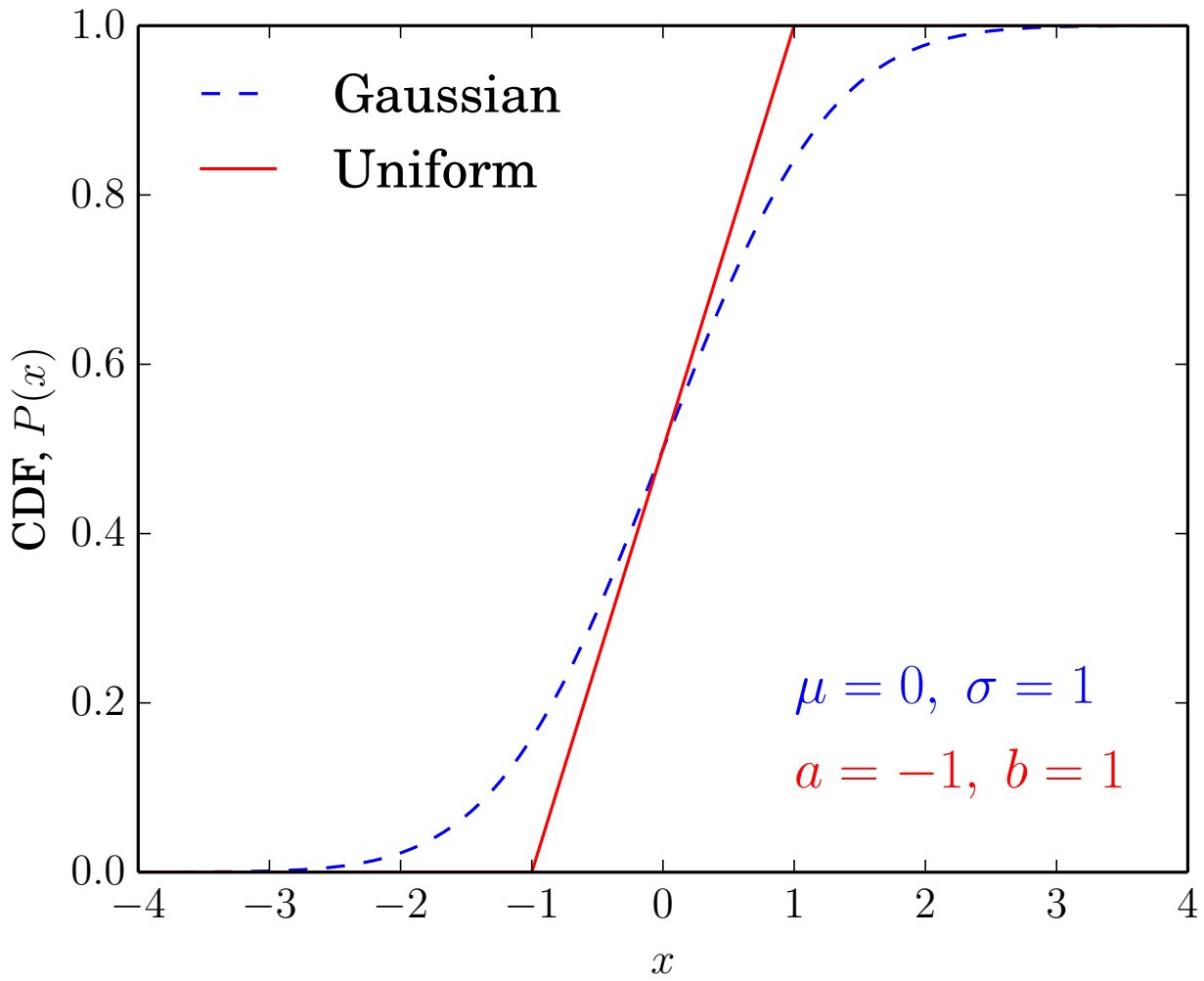
Cumulative distribution function (CDF)

Given a pdf, $p(x)$, the CDF defines the intergrated probability up to a given value z.

$$\boxed{P(z) = \int_{-\infty}^{z} p(x)\mathrm{d}x = \mathrm{Prob}(x < z)}$$

Note that the pdf is therefore the derivative of the CDF.

CUMULATIVE DISTRIBUTION FUNCTION (CDF)



The n^{th} **moment** of a pdf is defined as:

$$\langle x^n \rangle = \begin{cases} \sum_{x=0}^{\infty} x^n p(x|I) & \text{Discrete case} \\ \int_{-\infty}^{\infty} x^n p(x|I) \mathrm{d}x & \text{Continuous case} \end{cases}$$

The n^{th} **central moment**, where the origin is the mean, is defined by

$$\langle (x-\mu)^n \rangle = \begin{cases} \sum_{x=0}^{\infty} (x-\mu)^n p(x|I) & \text{Discrete case} \\ \int_{-\infty}^{\infty} (x-\mu)^n p(x|I) \mathrm{d}x & \text{Continuous case} \end{cases}$$

The 1^{st} moment is called the **expectation value** (or commonly the **mean**)

$$E(x) = \langle x \rangle = \begin{cases} \sum_{x=0}^{\infty} x p(x|I) & \text{Discrete case} \\ \int_{-\infty}^{\infty} x p(x|I) \mathrm{d}x & \text{Continuous case} \end{cases}$$

The 2^{nd} moment is called the **mean square**:

$$\langle x^2 \rangle = \begin{cases} \sum_{x=0}^{\infty} x^2 p(x|I) & \text{Discrete case} \\ \int_{-\infty}^{\infty} x^2 p(x|I) \mathrm{d}x & \text{Continuous case} \end{cases}$$

The **variance** (second *central* moment) is defined as:

$$\text{var}[x] = \sigma^2 = \langle (x-\mu)^2 \rangle = \begin{cases} \sum_{x=0}^{\infty} (x - \langle x \rangle)^2 p(x|I) & \text{Discrete case} \\ \int_{-\infty}^{\infty} (x - \langle x \rangle)^2 p(x|I) \mathrm{d}x & \text{Continuous case} \end{cases}$$

where σ is called the **standard deviation**.

In general the variance is given by:

$$\boxed{\text{var}[x] = \langle x^2 \rangle - \langle x \rangle^2}$$

The next two *central* moments are called **skewness** and **kurtosis**, which for a normal distribution measure the "lopsidedness" and width of the distribution's tails respectively.

The mean ($\langle x \rangle = \sum_{x=0}^{\infty} x p(x)$) of the Poisson pdf:

$$\begin{aligned}
\langle x \rangle &= \sum_{x=0}^{\infty} x \frac{\mu^x e^{-\mu}}{x!} \\
&= e^{-\mu} \sum_{x=0}^{\infty} x \frac{\mu^x}{x!} \\
&= e^{-\mu} \mu \underbrace{\sum_{x=1}^{\infty} \frac{\mu^{x-1}}{(x-1)!}}_{= e^{\mu} \text{ via } e^{y} = \sum_{n=0}^{\infty} \frac{y^n}{n!}} = \mu,
\end{aligned}$$

MOMENTS AND MEASURES OF A PDF

	$p(x)$	mean ($\langle x \rangle$)	variance
Poisson	$\frac{\mu^x e^{-\mu}}{x!}$	μ	μ
Binomial	$\frac{N!}{x!(N-x)!}\theta^x(1-\theta)^{N-x}$	$N\theta$	$N\theta(1-\theta)$
Uniform	$\frac{1}{b-a}$	$\frac{1}{2}(a+b)$	$\frac{1}{12}(b-a)^2$
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{(x-\mu)^2}{2\sigma^2}\right]$	μ	σ^2

The **median** value divides the CDF into two equal halfs:

$$P(x_{\text{median}}) = \int_{-\infty}^{x_{\text{median}}} p(x)\mathrm{d}x = 0.5.$$

The **mode** is the value of x for which the pdf is a **maximum** (modes of pdfs may not be uniquely defined, i.e. they can have more than one mode), i.e.

$$\frac{\partial p(x = x_{\text{mode}})}{\partial x} = 0$$

For a normal pdf we have mean $=$ median $=$ mode $= \mu$.

The variance of an arbitrary function of some random variable x can be approximated by

$$\boxed{\text{var}[f(x)] = \text{var}[x] \left(\frac{\partial f}{\partial x} \right)^2_{x=\bar{x}}}$$

where $\bar{x}$ is the mean of x. E.g. we measure variable x, but we want to determine the variance of $f(x) = y = x^2$, then we have

$$\text{var}[f(x)] = 4\sigma_x^2 \bar{x}^2.$$

If the pdf of the variable x is $p(x|I)$ and we have some function of x, $y = f(x)$ then what is the pdf, $p(y|I)$, of that function?

We require that probabilities over equivalent intervals are equal, so provided there is a one-to-one mapping between x and y we have

$$\int_{x_1}^{x_2} p(x|I)\mathrm{d}x = \int_{y_1=f(x_1)}^{y_2=f(x_2)} p(y|I)\mathrm{d}y,$$

and

$$p(y|I) = \underbrace{\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|}_{\text{Jacobian}} p(x|I)$$

More generally, if there is not a unique mapping between x and y we would have

$$p(y|I) = \sum_{f(x_i)=y} \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|_{x_i} p(x_i|I)$$

This can be expended to the multivariate case where M parameters $\{X_i\}$ map to M parameters $\{Y_i\}$ with

$$p(\{X_i\}|I) = p(\{Y_i\}|I) \times \underbrace{\left|\frac{\partial(Y_1, Y_2, \ldots, Y_M)}{\partial(X_1, X_2, \ldots, X_M)}\right|}_{\text{Jacobian}}$$

If there relation is between functions with different numbers of variables then change of variables may still be possible using dummy variables.

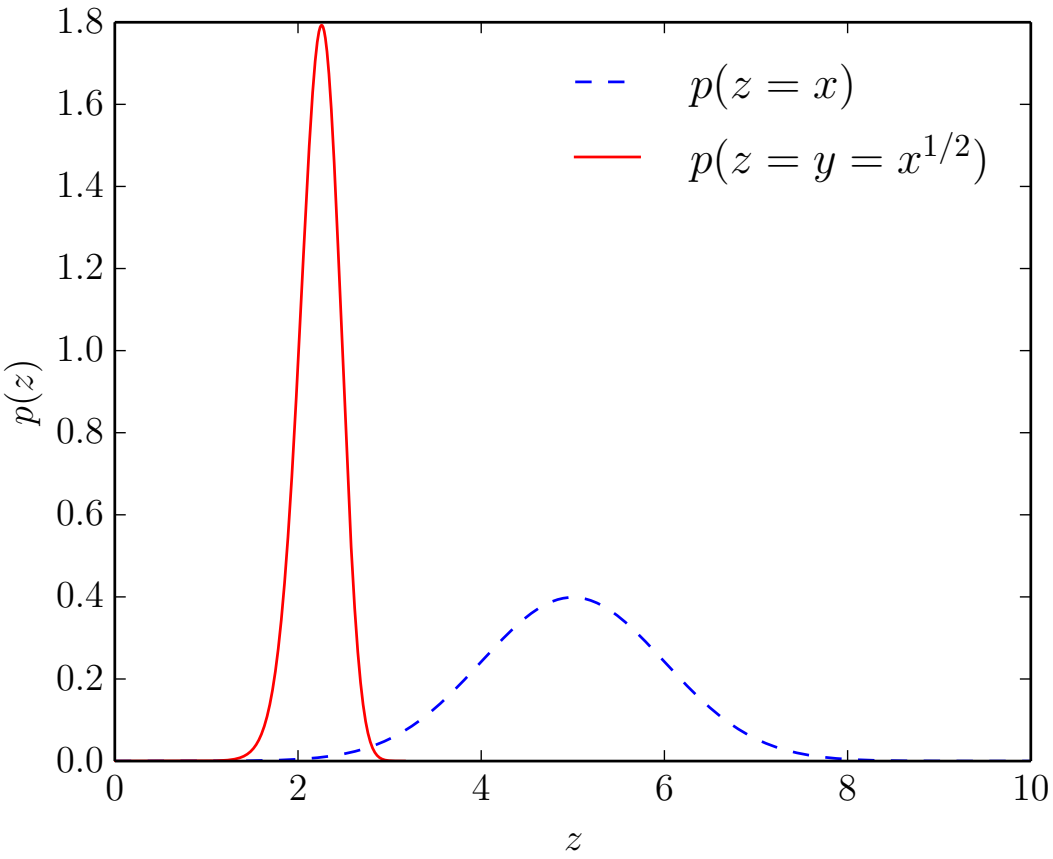
CHANGE OF VARIABLES: EXAMPLE

Suppose we have $y = \sqrt{x}$, then

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{2x^{1/2}},$$

and

$$\begin{aligned} p(y|I) &= \left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|^{-1} p(x|I) \\ &= 2x^{1/2} p(x|I). \end{aligned}$$



We have so far considered pdfs of single variables (univariate), but we can extend these to the case of two or more RVs (**multivariate**).

The **joint pdf** of two variables x and y is $p(x, y|I)$, where

$$\text{Prob}(a_1 < x < b_1 \text{ and } a_2 < y < b_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} p(x, y|I) \mathrm{d}x \mathrm{d}y.$$

This extends to any number of variables.

Given a **joint** pdf the marginal pdf on, say, x is given by:

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

which is a pdf in the sense that

- $p(x|I) \geq 0$ for all x
- Prob$(a < x < b) = \int_a^b p(x|I)\mathrm{d}x$
- $\int_{-\infty}^{\infty} p(x|I)\mathrm{d}x = 1$

Note that if the allowed range of a variable is known then the integral limits can be over that range, i.e. equivalent to saying that the probability is zero outside that range.

Given any multivariate pdf we can find the marginal pdf on any subset of the variables, e.g.

$$p(\theta, \phi|I) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(\theta, \phi, \psi, \lambda|I) \mathrm{d}\psi \mathrm{d}\lambda$$

Suppose we have two variables x and y where y is specified. We can write the **conditional** pdf for x *given* the value of y is true as $p(x|y, I)$ (we have already been implicitly using pdfs conditional on I).

From the **product rule** we can relate the *joint* pdf to the *conditional* pdf via

$$p(x|y, I) = \frac{p(x, y|I)}{p(y|I)} \text{ and similarly } p(y|x, I) = \frac{p(x, y|I)}{p(x|I)}.$$

Re-arranging these two equations leads back to **Bayes' theorem**

$$p(x|y, I) = \frac{p(y|x, I)}{p(y|I)} p(x|I)$$

If the conditional pdf $p(x|y, I)$ does not depend of y this means that x and y are statistically independent, i.e. the observed value of x is unaffected by the observed value of y.

So, for independent variables we have e.g. $p(x|y, I) = p(x|I)$ and $p(y|x, I) = p(y|I)$, so the joint pdf can be written

$$p(x, y|I) = p(x|y, I)p(y|I) = p(x|I)p(y|I)$$

We can say that a set of variables are **mutually independent** if their joint pdf can be written as the product of their marginal pdfs, i.e. n variables x_n are independent if

$$p(x_1, x_2, \ldots, x_n|I) = p(x_1|I)p(x_2|I) \ldots p(x_n|I).$$

A commonly used joint pdf for two variables x and y is the **bivariate normal** pdf, which has form

$$p(x, y|I) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}Q(x, y)\right]$$

where $Q(x, y)$ is a quadratic given by

$$Q(x, y) = \left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right).$$

This is defined by the parameters μ_x, μ_y, σ_x, σ_y and ρ. [This can be extended to a **multivariate normal** pdf.]

The first four parameters are given by:

$$\mu_x = E[x]$$
$$\mu_y = E[y]$$
$$\sigma_x^2 = \text{var}[x]$$
$$\sigma_y^2 = \text{var}[y]$$

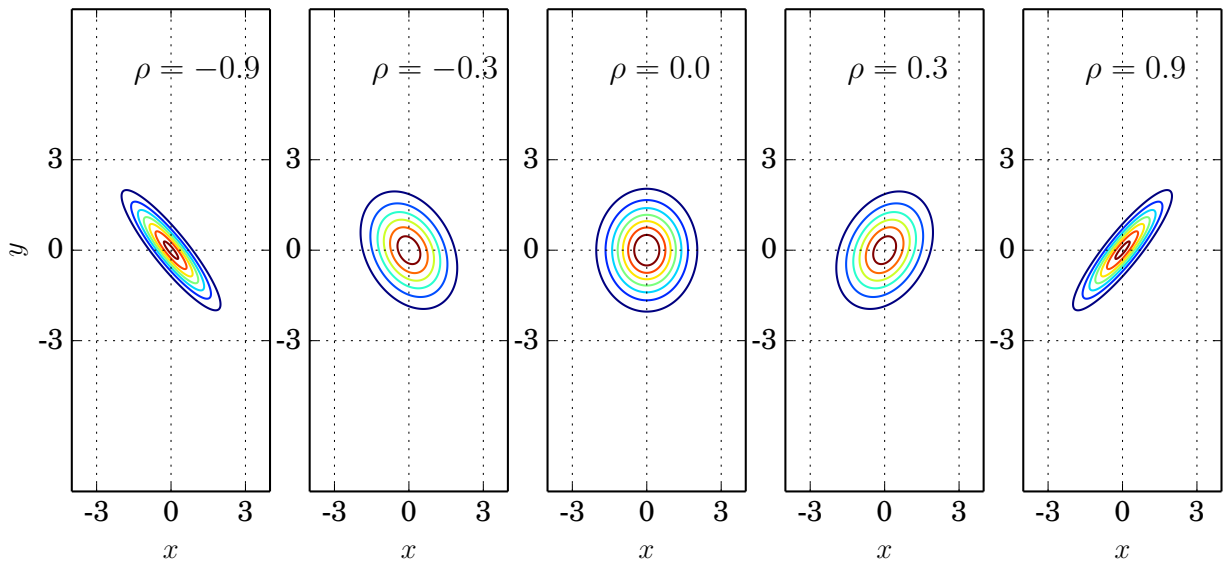
whilst the fifth, ρ, is known as the **correlation coefficient** and satisfies $\rho\sigma_x\sigma_y = E[(x-\mu_x)(y-\mu_y)]$

- $\rho = 0$ then x and y are independent.
- $\rho > 0$ gives a positive correlation - y tends to increase as x increases
- $\rho < 0$ gives a negative correlation - y tends to decrease as x increases

$E[(x-\mu_x)(y-\mu_y)]$ is known as the **covariance** of x and y and is often denoted by $\text{cov}(x, y)$.

Aside: Generally for any two variables with any pdf we define their covariance as

$$\text{cov}(x, y) = E[(x - E[x])(y - E[y])].$$



The **marginal** pdfs of x and y for the bivariate normal pdf are just the univariate normal pdfs

$$p(x|I) = N(\mu_x, \sigma_x^2)$$
$$p(y|I) = N(\mu_y, \sigma_y^2)$$

The **conditional** pdf of y given x is also a univariate normal pdf, but with

$$p(y|x, I) = N\left(\mu_y + \frac{\sigma_y}{\sigma_x}\rho(x - \mu_x), \sigma_y^2(1 - \rho^2)\right),$$

so as $|\rho| \to 1$ it can be seen that the width of the conditional distribution tends to zero.

$\mu_y + \frac{\sigma_y}{\sigma_x}\rho(x - \mu_x)$ is often referred to as the **conditional expectation** (value) of y given x, and the equation

$$y = \mu_y + \frac{\sigma_y}{\sigma_x}\rho(x - \mu_x)$$

is called the **regression line** of y on x.

BIBLIOGRAPHY

[1] R. T. Cox. *Am. J. Phys.*, 14(1), 1946.

[2] R. T. Cox. *The Algebra of Probable Inference*. Johns Hopkins University Press, 1961.