

# **Practical Statistics for Physics & Astronomy**

Ben Metcalf (robertbenton.metcalf@unibo.it)

**Lecture notes**

**<https://rbmetcalf.github.io/Practical-Statistics/lecturenotes/notes.pdf>**

**there is also a link to it on virtuale**

**Lab assignments will be available through the virtuale website**

# Practical Statistics for Astrophysics & Astronomy

## Laboratory:

- learn basic Python programming
- learn to use the statistical libraries in Python
- learn to read and manipulate data in Python
- learn to visualise data in Python
- learn to fit linear and nonlinear curves to data
- learn to make Bayesian & frequentist confidence level plots for multiple model parameters
- learn to do Monte Carlo calculations
- learn to do and plot Markov Chain Monte Carlo (MCMC) calculations and related methods
- learn to do supervised and unsupervised classification problems
- ...

Lab assignments will be listed on the website and due by the beginning of class the next week.

# Text Books

- "Modern Statistical Methods for Astronomy", Feigelson, E. D. & Babu, G. J. , 2012, Cambridge University Press  
*Covers statistics and specific applications to astronomy with many examples written in R. I don't find it to be very strong on explaining the concepts. Could be a good reference.*
- "Statistics in Theory and Practice", Lupton, R., 1993, Princeton University Press  
*Short book written by an astronomer. Lots of problems with answers. Short on explanation.*
- "Bayesian Logical Data Analysis for the Physical Sciences", Gregory, P.C., 2006, Cambridge University Press.  
*A somewhat ideological book written by an astronomer that promotes Bayesian methods.*
- "Numerical Recipes 3rd Edition: The Art of Scientific Computing" , Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P., 2007 Cambridge University Press.  
*A very practical book about numerical methods which has some good sections on statistical tests, Discrete Fourier Transforms, minimization algorithms, etc. .*

# Text Books

- "Data Analysis: A Bayesian Tutorial", Sivia, D.S. and Skilling, J. ,2006, 2nd. Oxford University Press  
*An introduction from a purely Bayesian viewpoint.*
- "Probability & Statistics", DeGroot, M.H., 1986, 2nd ed., Addison-Wesley,  
*Classic old-school text on statistics. Not always very practical.*
- "Probability Theory - The Logic of Science", Jaynes, E.T., 2003, Cambridge University Press.  
*Long winded and cantankerous, but a very interesting treatment of the connection between logic, probability and statistical physics. Not a book to start with.*

# Some inference problems in astrophysics

case	measurements	issues	parameters
1	magnitudes, fluxes $\tilde{m}$	noise background subtraction	"real" magnitudes and fluxes $m$
2	magnitudes, fluxes, distances $\tilde{m}, \tilde{D}$	2 forms of noise background subtraction	absolute magnitude, luminosity $M$
3	magnitudes, fluxes, distances $\{\tilde{m}_i, \tilde{D}_i\}$	2 forms of noise, missing data , false detections, background subtraction, ... expected variations between samples	luminosity function ( a statistical property) $\phi(L)$
4	radial velocities of a star $\{\tilde{v}_r(t_i)\}$	noise theory connecting parameters parameter degeneracies periodicity	$M_{\text{star}}, M_{\text{planet}},$ orbit - energy, ellipticity, inclination, orientation
5	stars in the MW $\{\tilde{v}_r\}, \{\vec{v}_p\}, \{\vec{\theta}\}, \{\tilde{m}\}, \{\tilde{D}\}, \dots$	noise, incompleteness intermediate statistics	theoretical statistics $\sigma_*^2(\mathbf{x})$ $\rho_*(\mathbf{x})$ $M_{\text{bulge}}, \Sigma_{\text{disk}},$
6	galaxy redshift $\{\tilde{z}_{\text{phot}}\}, \{\tilde{\vec{\theta}}\}, \{\tilde{m}\}$	noise , incompleteness intermediate statistics	theoretical statistics $\sigma^2(R) = \left\langle \left( \frac{M(\mathbf{x}; R) - \bar{M}(R)}{\bar{M}(R)} \right)^2 \right\rangle$ or $P(\mathbf{k})$ or $\xi(r)$ $\Omega_m, \Omega_\Lambda, H_o, w, \dots$
7	radio visibilities $V(u, v)$ Image deconvolution $I(x, y)$	noise, incompleteness more parameters than measurements	image pixel values

Table 2.1: Some inference problems in astrophysics.

## Inference Problem

parameter estimation

regression - curve fitting

model fitting

“training” a model or machine

## Hypothesis testing

frequentist tests

confidence levels

## Model Selection

Are extra parameters justified

validation

detection

## Prediction

machine learning

## Experimental Design

predicting errors without data

information content of data

## Classification

dimensional reduction

supervised and unsupervised classification

# **Topics**

**Probability distributions**

**Bayesian statistics**

**Linear modes, least-squares and regression**

**Supervised learning & resampling techniques**

**Hypothesis testing & frequentist parameter fitting**

**Categorical variables**

**Maximum likelihood, Fisher information & experimental design**

**Numerical methods for inference**

**Information and Entropy**

**Random fields & power spectrum estimation**

**Artificial Neural networks and machine learning**

# Frequentist Interpretation of Probability

$$P(A) = \lim_{N \rightarrow \infty} \frac{\text{number of trials where A is true}}{N(\text{total number of trials})}$$

# **Classical "Interpretation" of Probability**

**Fundamental events or states.**

**Principle of Indifference :** Each of  $n$  mutually exclusive events that encompass all possibilities should be given probability  $1/n$  if there is no reason to favor one over any other.

**The probability of a macro-event or property is the sum of the probabilities of the micro-events or states that have that property.**

Count the states.

What are fundamental states? How can they be identified?

Can all problems be reduced to fundamental states/events in this way?

## Desiderata of Bayesian Probability theory (from Gregory 2006)

- 1 Degrees of plausibility are represented by real numbers.
- 2 The measure of plausibility must exhibit qualitative agreement with rationality. This means that as new information supporting the truth of a proposition is supplied, the number which represents the plausibility will increase continuously and monotonically. Also, to maintain rationality, the deductive limit (plausibility 1 and 0) must be obtained where appropriate.
- 3 Consistency
  - 1 *Structured consistency* : If the conclusion can be reasoned out in more than one way, every possible way must lead to the same result.  
( Logically equivalent statements must have the same weight. )
  - 2 *Propriety*: The theory must take account of all information that is relevant to the question.
  - 3 *Jaynes consistency*: Equivalent states of knowledge must be represented by equivalent plausibility assignments. For example, if  $(A, B) \parallel C = B \parallel C$ , then the plausibility of  $(A, B) \parallel C$  must equal the plausibility of  $B \parallel C$ . (Here  $\parallel$  is logical "or" and  $,$  is logical "and").

$A$	$B$	$A, B$	$\overline{A, B}$	$\overline{A} \parallel \overline{B}$	$A \parallel B$	$\overline{A \parallel B}$	$\overline{\overline{A}, \overline{B}}$	$\overline{\overline{A}, B}$	$\overline{A}, \overline{B}$	$\overline{A} \parallel B$	$\overline{A} \parallel \overline{B}$
F	T	F	T	T	T	F	F	T	F	T	F
F	F	F	T	T	F	T	T	F	F	T	T
T	T	T	F	F	T	F	F	F	F	T	T
T	F	F	T	T	T	F	F	F	T	F	T

Table 1.1: The truth table for binary logical expressions. Statements with the same truth table are logically equivalent. Note that  $\overline{A, B} = \overline{A} \parallel \overline{B}$  and  $\overline{A \parallel B} = \overline{A}, \overline{B}$  because their truth tables are the same.

# Fundamental rules of probability

$$P(A) \geq 0$$

$$P(A, B) = P(A)P(B|A)$$

$$P(A) + P(\overline{A}) = 1$$

**positive semidefinite  
product rule  
sum rule**

(1)

# Fundamental rules of probability

$$P(A) \geq 0$$

$$P(A, B) = P(A)P(B|A)$$

$$P(A) + P(\bar{A}) = 1$$

**positive semidefinite  
product rule  
sum rule**

(1)

$P(B, A)$  = probability of both  $A$  **and**  $B$  being true  
**joint probability**

$P(B|A)$  = probability of  $B$  **given that**  $A$  is true  
**conditional probability**

$P(A||B)$  = probability of  $A$  **or**  $B$  being true

$$\begin{aligned}
P(A||B) &= 1 - P(\overline{A}||\overline{B}) && \text{see table 1.1} \\
&= 1 - P(\overline{A}, \overline{B}) && \text{product rule} \\
&= 1 - P(\overline{A})P(\overline{B}|\overline{A}) && \text{sum rule} \\
&= 1 - P(\overline{A}) [1 - P(B|\overline{A})] && \\
&= 1 - P(\overline{A}) - P(\overline{A})P(B|\overline{A}) && \\
&= P(A) + P(\overline{A})P(B|\overline{A}) && \text{sum rule} \\
&= P(A) + P(\overline{A}, B) && \text{product rule} \\
&= P(A) + P(B)P(\overline{A}|B) && \text{product rule} \\
&= P(A) + P(B) [1 - P(A|B)] && \text{sum rule} \\
&= P(A) + P(B) - P(B)P(A|B) && \\
P(A||B) &= P(A) + P(B) - P(B, A) && \text{extended sum rule}
\end{aligned}$$

$$\begin{cases} P(A, B) = P(A)P(B|A) \\ P(A) + P(\bar{A}) = 1 \end{cases}$$

product rule  
sum rule

$$P(B)P(A|B) = P(A)P(B|A) \quad \text{Bayes' theorem}$$

$$P(A||B) = P(A) + P(B) - P(A, B) \quad \text{extended sum rule}$$

**independent events**

$$P(A|B) = P(A) \longrightarrow P(B|A) = P(B) \longrightarrow P(A, B) = P(A)P(B)$$

**mutually exclusive events**

$$P(A, B) = 0 \longrightarrow P(A||B) = P(A) + P(B)$$

**Example:**

## **Rolling Dice**

Probability of getting a **6 or a 5** :

$$\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Probability of getting a **6 and then a 5** :

$$\left(\frac{1}{6}\right) \left(\frac{1}{6}\right) = \frac{1}{36}$$

Probability of getting a **6 and a 5 in any order** :

$$2 \left(\frac{1}{6}\right) \left(\frac{1}{6}\right) = \frac{1}{18} \quad \left(\frac{1}{3}\right) \left(\frac{1}{6}\right) = \frac{1}{18}$$

**Problem 1.** We know the probability of a person having red hair is  $P(R)$ , the probability of a person having blue eyes is  $P(B)$  and that the probability of a red headed person having blue eyes is  $P(B|R)$ .

1. What is the probability that a blue eyed person will have red hair?
2. What is the probability that a person will have both blue eyes and red hair?
3. What is the probability that a person will have either blue eyes or red hair?

Observations  $\{A_i\}$  are all mutually exclusive and together they include all possible outcomes then

$$\begin{aligned}
 1 &= P(A_1||A_2||A_3||\dots|B) + P(\overline{A_1||A_2||A_3||\dots}|B) && \text{sum rule} \\
 &= P(A_1||A_2||A_3||\dots|B) + 0 && \text{can't all be false} \\
 &= P(A_1|B) + P(A_2||A_3||\dots|B) && \text{extended sum rule} \\
 &= P(A_1|B) + P(A_2|B) + P(A_3||\dots|B) && \text{extended sum rule} \\
 &= \sum_i P(A_i|B)
 \end{aligned}$$

The sum of the probabilities for all possible outcomes is one.

In the continuum limit

$$1 = \int_{-\infty}^{\infty} p(x|y)dx$$

# COVID test example

Say we have developed a new cheaper test COVID-19. We test it on patients that we know have COVID-19 and find that 90% of them get a positive result. Then we test it on patients that we know don't have COVID-19 and we find that 90% of them get a negative test result. From previous research we know that the COVID rate in the general population is 16 per 10,000.

If we use this test in the general population what is the chance of a person with a positive test of actually having COVID-19?

$C$  = has COVID

$T$  = tests positive

$\bar{C}$  = doesn't have COVID

$\bar{T}$  = tests negative

We have measurements of :

$P(T|C)$  ,  $P(\bar{T}|\bar{C})$  and  $P(C)$

# COVID test example

If we use this test in the general population what is the chance of a person with a positive test of actually having cancer?

$C$  = has COVID

$T$  = tests positive

$\bar{C}$  = doesn't have COVID

$\bar{T}$  = tests negative

$$P(C|T) = \frac{P(T|C)P(C)}{P(T)}$$

Bayes' theorem

$$\begin{aligned} P(T) &= P(T|\bar{C})P(\bar{C}) + P(T|C)P(C) = P(T|\bar{C})[1 - P(C)] + P(T|C)P(C) \\ &= [1 - P(\bar{T}|\bar{C})][1 - P(C)] + P(T|C)P(C) \\ &= [1 - P(\bar{T}|\bar{C})] + [P(\bar{T}|\bar{C}) + P(T|C) - 1]P(C) \end{aligned}$$

$$\begin{aligned} P(C|T) &= \frac{P(T|C)P(C)}{[1 - P(\bar{T}|\bar{C})] + [P(\bar{T}|\bar{C}) + P(T|C) - 1]P(C)} \\ &= 0.014 \end{aligned}$$

# Marginal distribution

$A_i$     mutually exclusive events (only one can be true at a time)

$$P(B) = \sum_{\forall i} P(B|A_i)P(A_i) = \sum_{\forall i} P(B, A_i) = \sum_{\forall i} P(A_i|B)P(B) = P(B) \sum_{\forall i} P(A_i|B)$$

continuous case

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

$p(x)$     probability density function or probability mass function

## **flip a coin**

Probability of getting a particular sequence of heads and tails :

$$P(1, 2, 2, 1, 2) = p_1 p_2 p_2 p_1 p_2 = (p_1)^2 (p_2)^3 = (p_1)^2 (1 - p_1)^3$$

Probability of getting a particular number of heads and tails in any order :

All orderings have the same probability so we just need to know how many orderings there are

$$P(\{1, 2, 2, 1, 2\}) = \frac{5!}{2!(5-2)!} (p_1)^2 (1 - p_1)^3 = \binom{5}{2} (p_1)^2 (1 - p_1)^3$$

1,2,3

$$N^n$$

Number of combinations of  $n$  objects taken from  $N$  possibilities with replacement

(1,1) (1,2) (1,3)  
(2,1) (2,2) (2,3)  
(3,1) (3,2) (3,3)

$$N!$$

Number of permutations of  $n$  objects

(1,2,3) (1,3,2) (2,1,3) (2,3,1)  
(3,2,1) (3,1,2)

$$N(N - 1)(N - 2)...(N - n + 1) = \frac{N(N - 1)(N - 2)...1}{(N - n)(N - n - 1)...1} = \frac{N!}{(N - n)!}$$

(1,2) (1,3) (2,1) (2,3) (3,2) (3,1)

Number of *ordered* combinations of  $n$  objects taken from  $N$  *without replacement*.

$$\frac{N!}{n!(N - n)!} = \binom{N}{n}$$

Number of *unordered* sets of  $n$  objects taken from  $N$  *without replacement*  
“ $N$  choose  $n$ ”

(1,2) (1,3) (2,3)

$$\binom{n + N - 1}{n}$$

Number of *unordered* sets of  $n$  objects taken from  $N$  *with replacement*

(1,1) (1,2) (1,3)  
(2,2) (2,3)  
(3,3)

# Binomial Distribution

$$P(n|N,p) = \binom{N}{n} p^n (1-p)^{N-n}$$

$$Var[n] = \langle n^2 \rangle - \langle n \rangle^2$$

$$\begin{aligned} \langle n \rangle &= \sum_{n=0}^{\infty} n p(n|N) = Np \\ &= \sum_{n=0}^{\infty} n^2 p(n|N) - \langle n \rangle^2 \\ &= Np(1-p) \end{aligned}$$

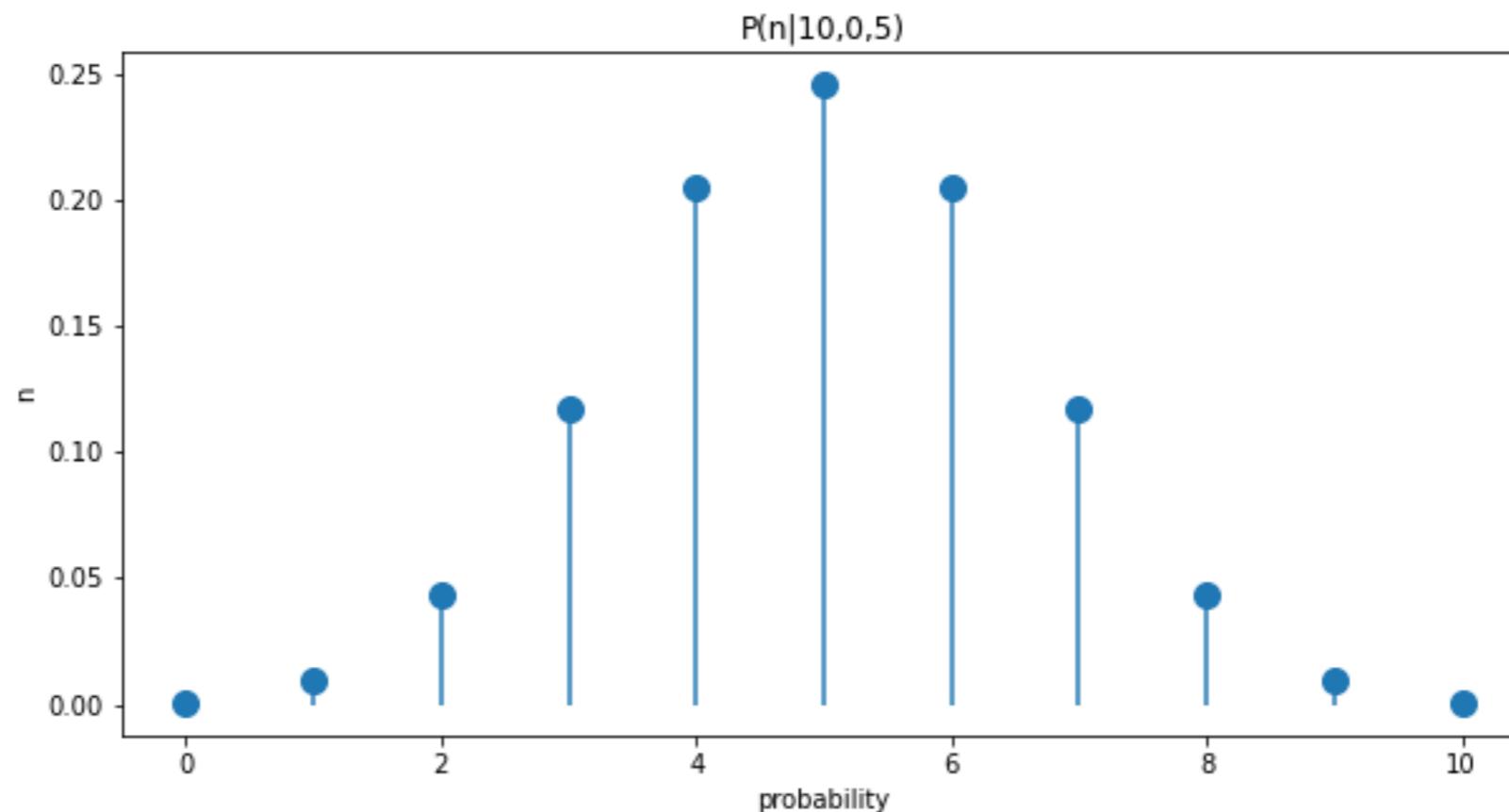
# Binomial Distribution

$$P(n|N, p) = \binom{N}{n} p^n (1 - p)^{N-n}$$

**coin flips**

Chance of getting 7 heads out of 10 flips:

$$P(7|10, p) = \binom{10}{7} p^7 (1 - p)^3 \quad P(7|10, 0.5) = 0.117$$

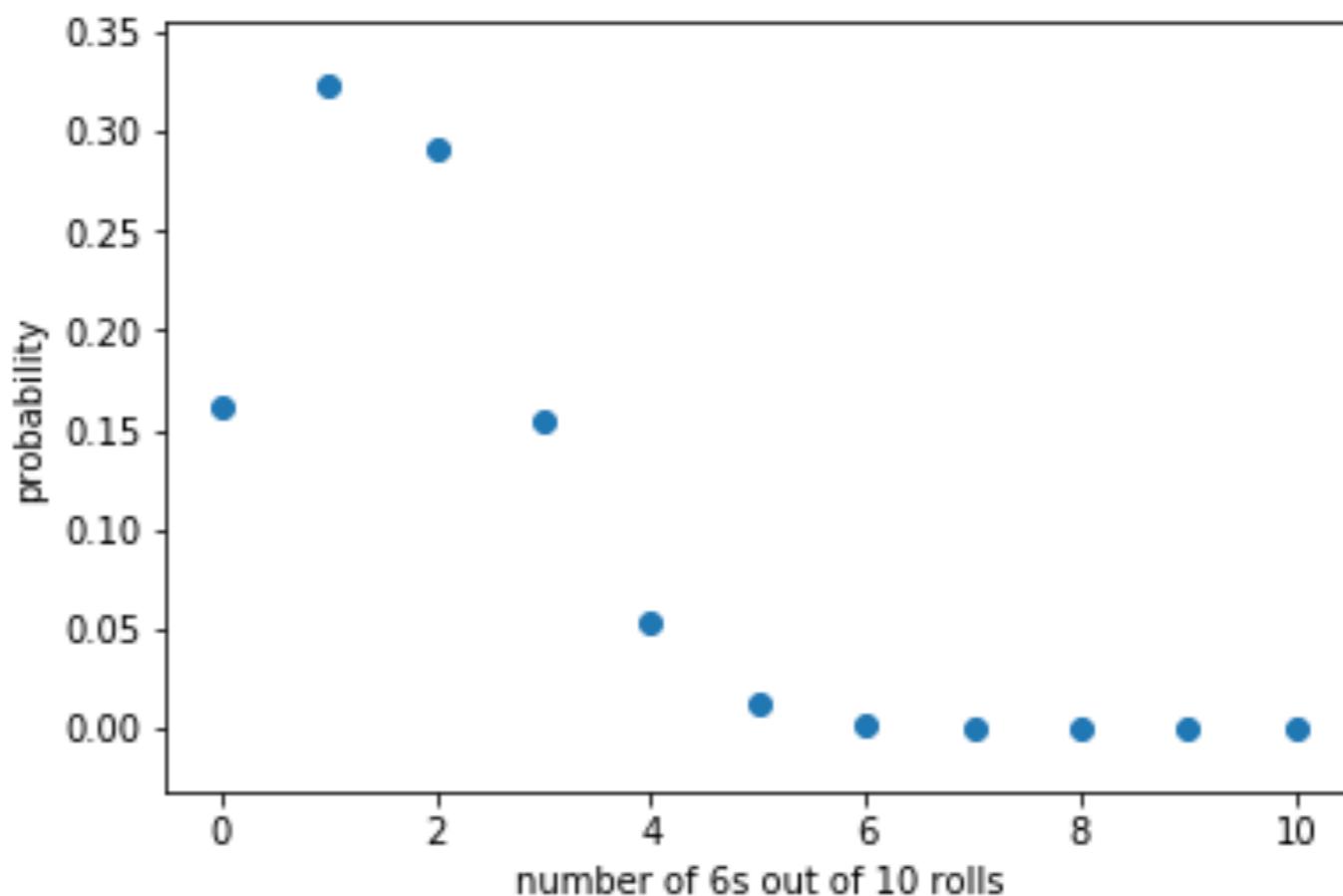


# Binomial Distribution

$$P(n|N,p) = \binom{N}{n} p^n (1-p)^{N-n}$$

**rolling dice**

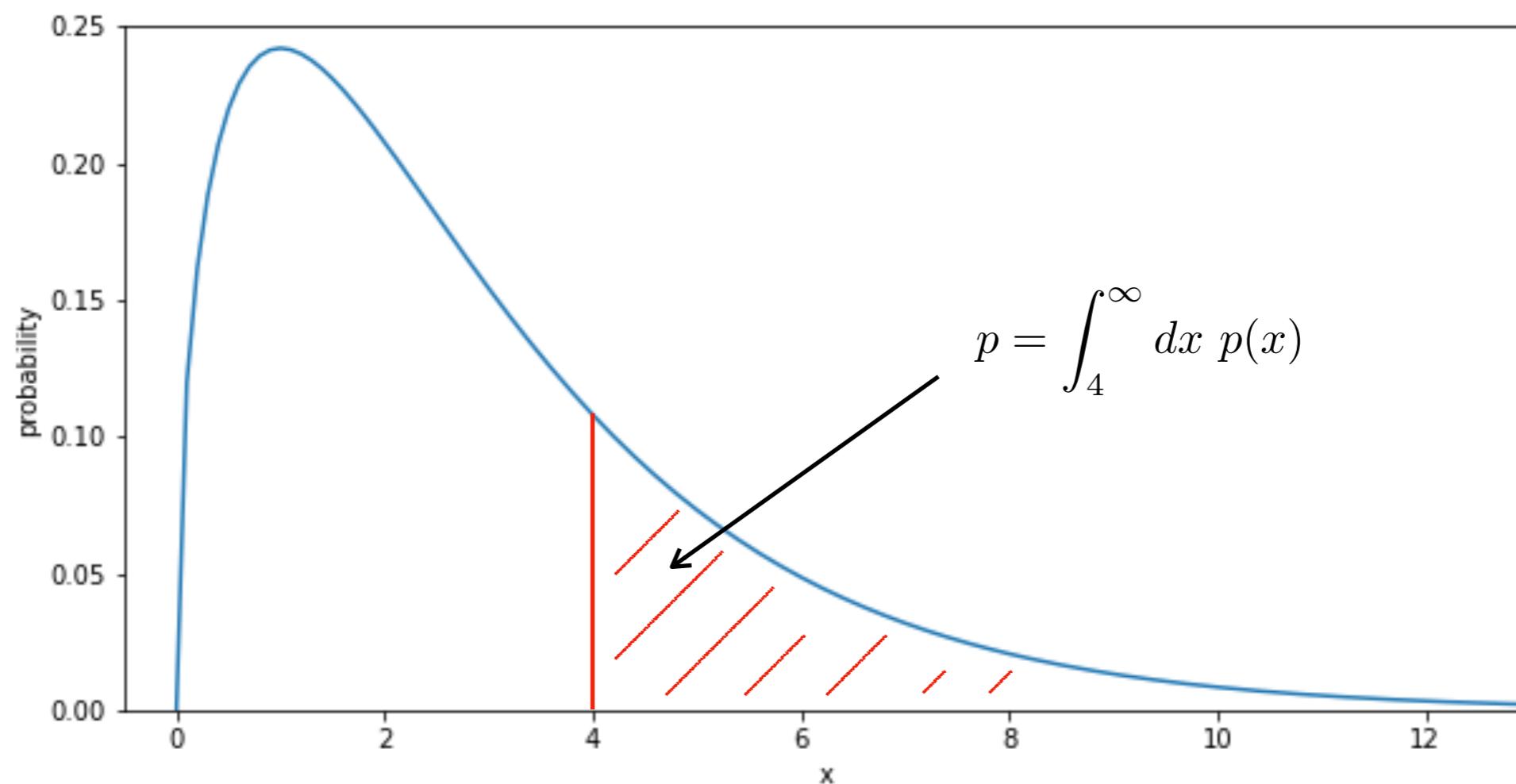
Number of 6s out of 10 rolls of a fair die       $p = 1/6$



# Binomial Distribution

$$P(n|N,p) = \binom{N}{n} p^n (1-p)^{N-n}$$

If you take N samples from a distribution, the number in any range is binomially distributed.



## Rolling unfair dice

Probability of getting specific numbers in a specific order

$$P(1, 2, 2, 5, 5) = p_1 p_2 p_2 p_5 p_5 = (p_1)(p_2)^2(p_5)^2$$

Probability of getting specific numbers in any order

$$P(\{1, 2, 2, 5, 5\}) = \frac{5!}{1!2!2!} (p_1)(p_2)^2(p_5)^2$$

Rewriting this as

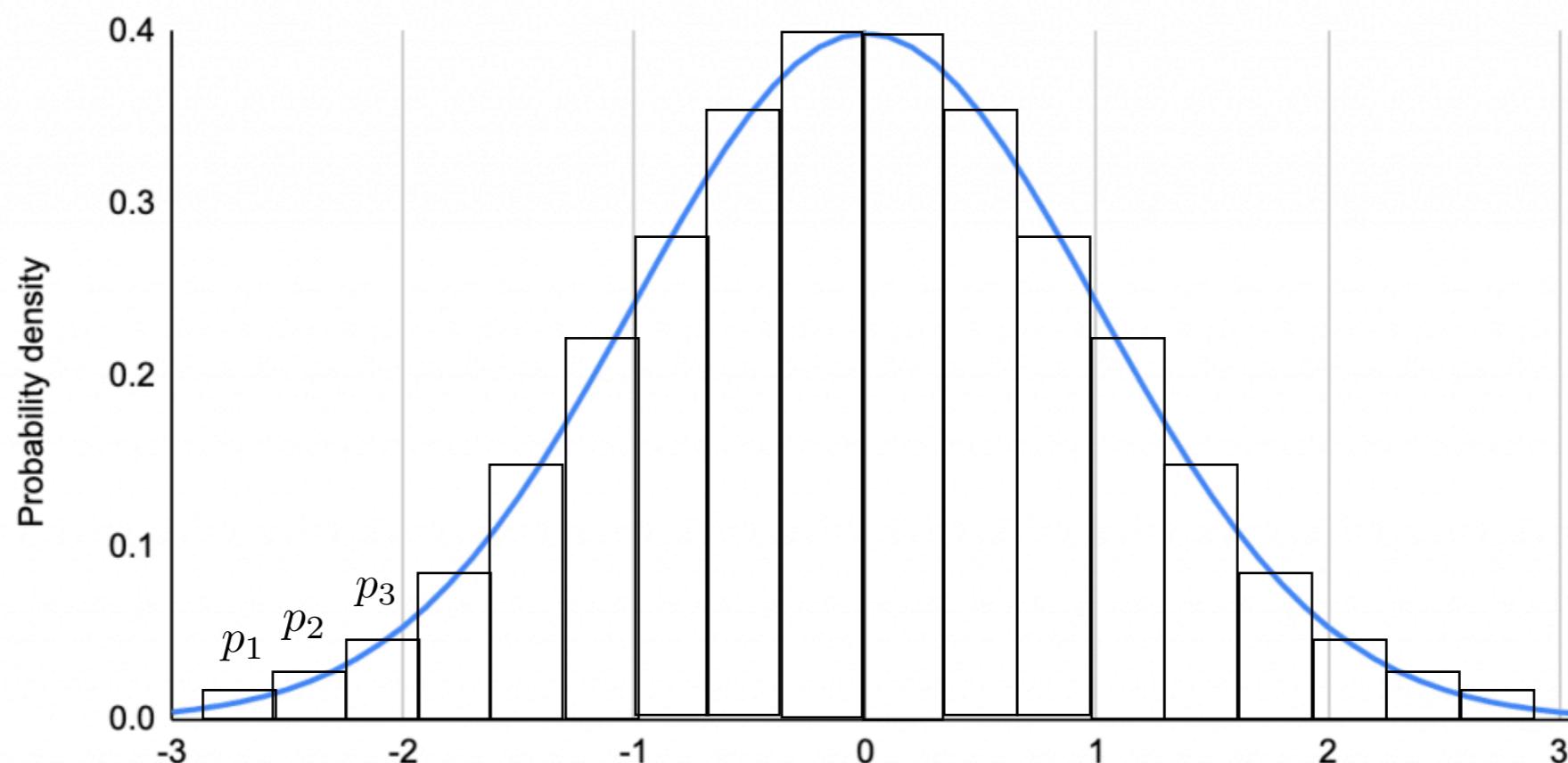
$$P(n_1, n_2, n_3, n_4, n_5, n_6 | N, \{p_i\}) = P(1, 2, 0, 0, 2, 0 | N, \{p_i\})$$

## Multinomial Distribution

$$\begin{aligned} P(n_1, n_2, \dots, n_k | N, \{p_i\}) &= \frac{N!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} & \sum_{i=0}^k p_i = 1 \\ &= \frac{N!}{\prod_{i=1}^k n_i!} \prod_{i=1}^k p_i^{n_i} & \sum_{i=0}^N n_i = N \end{aligned}$$

# Multinomial Distribution

$$\begin{aligned} P(n_1, n_2, \dots, n_k | N, \{p_i\}) &= \frac{N!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} \\ &= \frac{N!}{\prod_{i=1}^k n_i!} \prod_{i=1}^k p_i^{n_i} \end{aligned}$$



- 1) come up with a team name
- 2) create a CSV file *yourteamname.csv*
  - one column must be called "Flip"
  - in this column 0 for heads, 1 for tails
  - 200 trials
- 3) email it to me at [robertbenton.metcalf@unibo.it](mailto:robertbenton.metcalf@unibo.it)

## Cumulative Distribution Function (CDF)

$$F(x) = \int_{-\infty}^x dx' p(x')$$

## Empirical Cumulative Distribution Function (ECD)

$$\hat{F}(x) = \frac{N(< x)}{N} = \frac{1}{N} \sum_{i=1}^N \Theta(x_i < x)$$

$$\Theta(A) = \begin{cases} 1 & , \quad A \text{ is true} \\ 0 & , \quad A \text{ is false} \end{cases}$$

$N(< x)$     number of sample below x



## Birthday "Paradox"

What is the probability that two people out of  $n$  will have the same birthday?

This must be 1 minus the probability that no one has the same birthday.

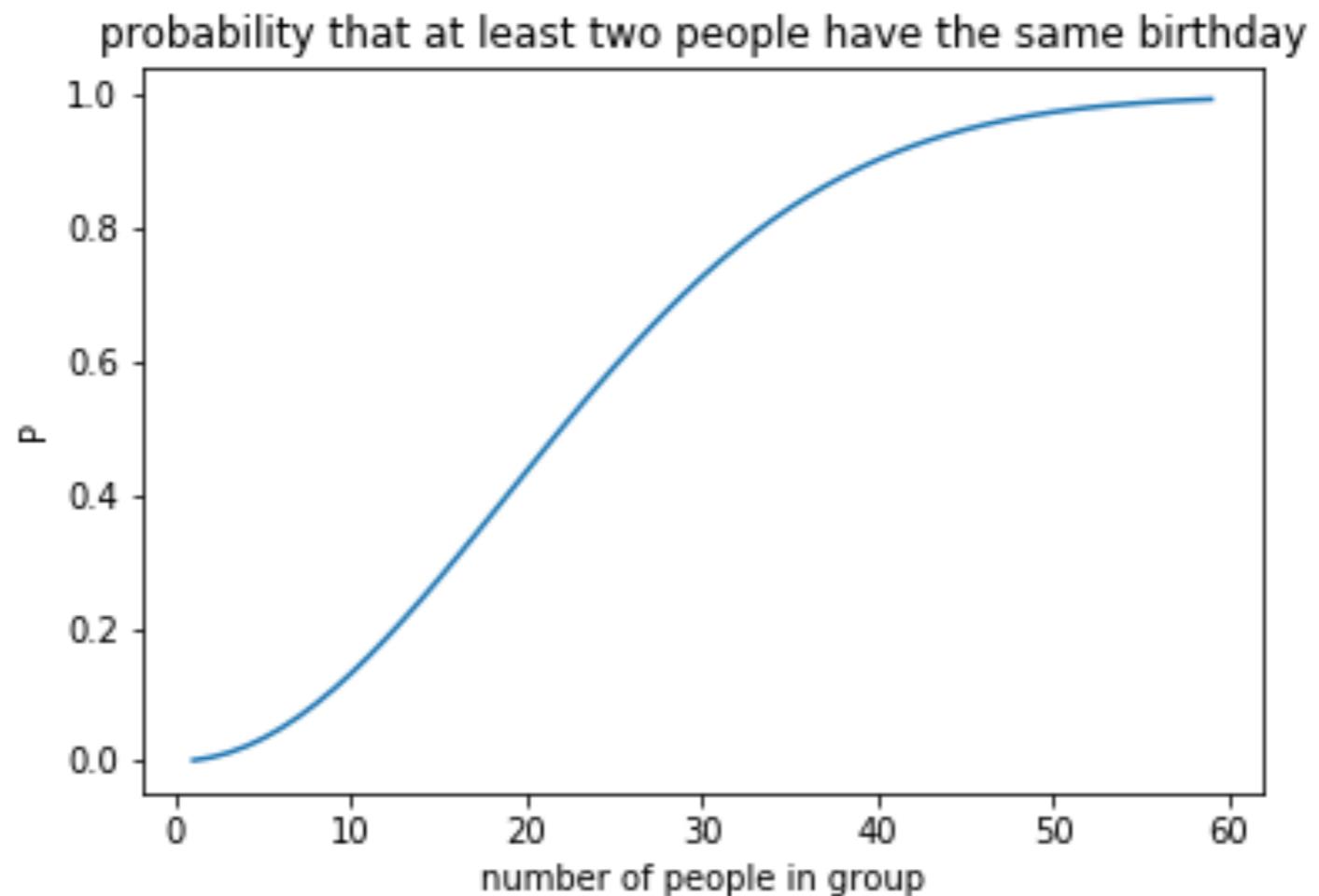
There are  $365^n$  possible combinations of birthdays (no leap days).

How many ways can we pick  $n$  birthdays without repeating any?

$$365 \times 364 \times \dots \times (365 - n + 1) = 365! / (365 - n)!$$

$$P(\text{at least two the same}) = 1 - P(\text{no two the same}) = 1 - \frac{365!}{365^n (365 - n)!}.$$

$$\ln N! \simeq N \ln N - N \quad \text{Stirling's approximation}$$



$$P(N) \simeq 1 - \left( \frac{N}{N-n} \right)^{N-n} e^{-n} \quad N = 356$$

- **cumulative distribution function** - the function of  $x$  describing the probability of the measured value being  $< x$ :

$$F(x) = \int_{-\infty}^x dx' p(x') \quad (13)$$

By definition  $F(-\infty) = 0$  and  $F(+\infty) = 1$ . The cumulative distribution for a discrete distribution is defined in the obvious way.

- **Quantile function** is the inverse of the cumulative distribution function

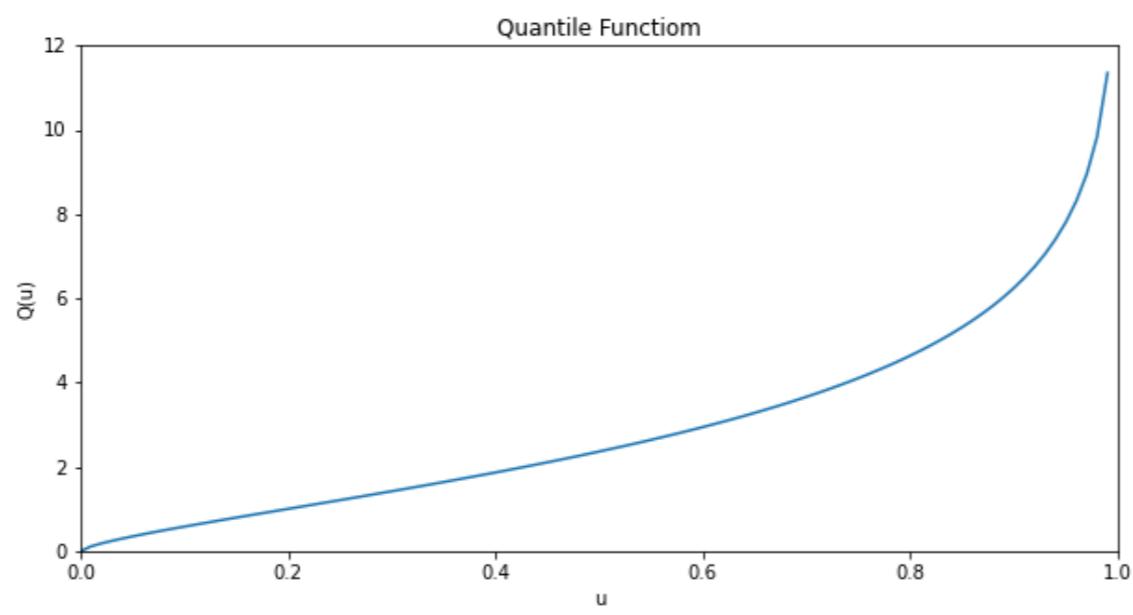
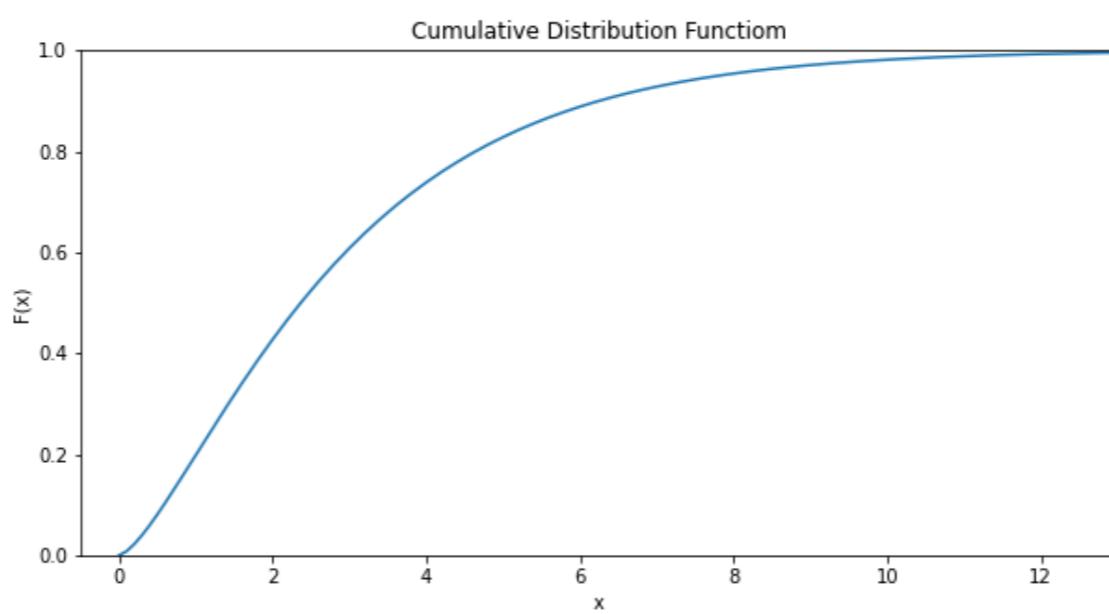
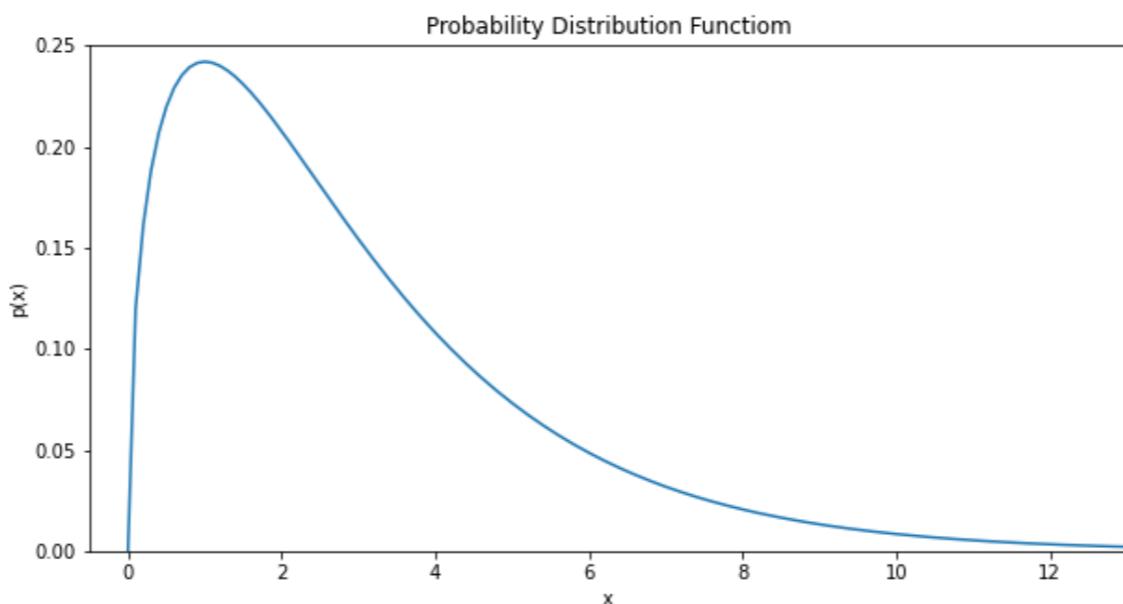
$$Q(u) = F^{-1}(u) \quad 0 \leq u \leq 1 \quad (14)$$

There is a probability  $u$  that the random variable will be  $x < Q(u)$ .

- **expectation value** - The "average" of any function of the random variable. This is denoted by  $E[\dots]$  or  $\langle \dots \rangle$ . The expectation value of  $f(x)$  is

$$E [f(x)] = \langle f(x) \rangle = \left\{ \begin{array}{l} \sum_x p(x) f(x) \\ \int_{-\infty}^{\infty} dx p(x) f(x) \end{array} \right. \quad (15)$$

- **mode** - A point where a distribution has a maximum. **Unimodal** distributions have one mode and **multimodal** distributions have more than one.



- **median** -  $Q(1/2)$  or the point in the distribution where  $F(x) = 1/2$ .  
The probability that  $x$  will be less than the median is equal to the probability that it will be more than the median. In a sample or data set the median is the data point that has equal numbers of data points larger than and less than it. For a set with an even number of points the arithmetic mean between the two points closest to having this property is often used.
- **mean** - The mean is the expectation value of the random variable itself,  $E[x]$ . This will often be represented by  $\mu$ .
- **moments** - The  $n$ th moment of a distribution is  $E[x^n]$ .
- **central moments** - The  $n$ th central moment is  $E[(x - \mu)^n]$
- **variance** - The variance is the second central moment  $E[(x - \mu)^2]$ . It is often denoted by  $Var[x]$  or  $\sigma^2$ . This is a measure of the width of the distribution.
- **standard deviation** - the square root of the variance. It is often denoted by  $\sigma$ . An equivalent measure of the width of the distribution in the same units as the random variable.

- **mean deviation**  $E[|x - \mu|]$ . This is an alternative measure of the width of a distribution. It is often more robustly estimated from a small sample especially when the distribution has large "tails" (much of the probability lies far away from the peak or beyond  $\sim \sigma$  from it.).
- **skewness** -  $E[(x - \mu)^3]/\sigma^3$ . This is a unitless measure of the asymmetry of the distribution.
- **kurtosis** -  $E[(x - \mu)^4]/\sigma^4$ . This is a measure of the relative importance of outliers (point differing from the mean by larger than several  $\sigma$ ). If the kurtosis is larger than 1 the "tails" of the distribution are more important than for a Gaussian. This also reflects the "boxyness" of the distribution.
- **standardized variable** - It is often useful to rescale a random variable with the standard deviation and mean of its distribution

$$X = \frac{(x - \mu)}{\sigma}. \quad (16)$$

This variable will always have a mean of 0 and a variance of 1.

The characteristic function is essentially the same thing as the moment generating function only it is the Fourier transform of the probability distribution instead of the Laplace transform

$$\phi_x(t) = \langle e^{itx} \rangle = \left\{ \begin{array}{l} \sum_x e^{itx} p(x) \\ \int_{-\infty}^{+\infty} dx e^{itx} p(x) \end{array} \right. \quad (17)$$

The only difference is the  $i$ .

The characteristic function has the following properties under transformations of the random variable

$$\begin{aligned} \phi_z(t) &= \phi_x(t/n) & z &= x/n \\ \phi_z(t) &= e^{-it\mu/\sigma} \phi_x(t/\sigma) & z &= \frac{x-\mu}{\sigma} \\ \phi_z(t) &= \phi_x(t)\phi_y(t) & z &= x + y \end{aligned} \quad (18)$$

where  $x$  and  $y$  are independently distributed. They are readily derived and recognizable as properties of the Fourier transform. The last one comes from the convolution theorem.

## Multivariant distribution

$$p(x_1, x_2, \dots, x_k) dx_1 dx_2 \dots dx_k \quad (42)$$

expectation value of an arbitrary function  $f(x_1, x_2, \dots, x_k)$

$$E[f(x_1, x_2, \dots, x_k)] = \int \dots \int dx_1 \dots dx_k f(x_1, x_2, \dots, x_k) p(x_1, x_2, \dots, x_k) \quad (43)$$

$$= \prod_{i=1}^k \int dx_i f(x_1, x_2, \dots, x_k) p(x_1, x_2, \dots, x_k) \quad (44)$$

This is also written  $\langle f(x_1, x_2, \dots, x_k) \rangle$  or  $\overline{f(x_1, x_2, \dots, x_k)}$ . The probability distribution is normalized so  $E[1] = 1$ .

# covariance matrix

$$C_{ij} = Cov[x_i x_j] \equiv E[(x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle)]$$

$$C_{xy} = E[xy] - \bar{x}\bar{y}$$

## Cauchy–Schwarz inequality

$$C_{xy}^2 = |Cov[x, y]|^2 \leq Var[x]Var[y]$$

- $\mathbf{C}$  is symmetric,  $C_{ij} = C_{ji}$ .
- $C_{ii} \geq 0$  for all  $i$
- the eigenvalues are  $\geq 0$
- $\mathbf{x}^T \mathbf{C} \mathbf{x} \geq 0$  for all non zero  $\mathbf{x}$ , i.e.  $\mathbf{C}$  is a positive semi-definite matrix
- $\mathbf{C} = \mathbf{M} \Lambda \mathbf{M}^T$  where  $\Lambda$  is diagonal and  $\mathbf{M}$  is an orthogonal matrix  
 $\mathbf{M}^{-1} = \mathbf{M}^T$ ,  $|\mathbf{M}| = 1$ .

- $\mathbf{C}^{-1} = \mathbf{M}\Lambda^{-1}\mathbf{M}^T$
- $\mathbf{C}^{-1}$  is the *precision matrix*
- $\rho$  The *correlation matrix*

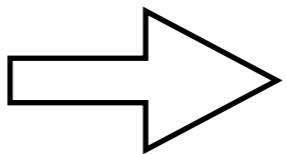
$$\rho_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}} = \frac{C_{ij}}{\sigma_{ii}\sigma_{jj}}$$

$$\rho = \Lambda^{-1/2} \mathbf{C} \Lambda^{-1/2}$$

- Principle Components  $\mathbf{y} = \mathbf{M}^T(\mathbf{x} - \boldsymbol{\mu})$

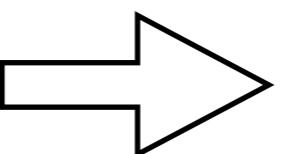
$$\langle \mathbf{y} \mathbf{y}^T \rangle = \text{Cov}[\mathbf{y}, \mathbf{y}] = \Lambda$$

Independent



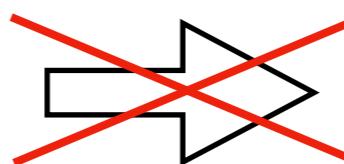
uncorrelated

correlated



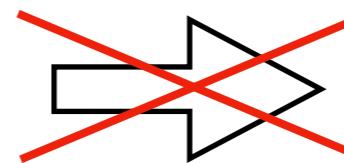
dependent

uncorrelated



Independent

dependent



correlated

# Principle Components

$$\mathbf{y} = \mathbf{M}^T(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{M}^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

Where the columns of  $\mathbf{M}$  are the normalised eigenvectors of the covariance matrix  $\mathbf{C}$

Because  $\mathbf{M}$  is orthogonal  $\mathbf{M}^T = \mathbf{M}^{-1}$

$\langle \mathbf{y} \mathbf{y}^T \rangle = \boldsymbol{\Lambda}$  where  $\boldsymbol{\Lambda}$  is diagonal with the eigenvalues of  $\mathbf{C}$  for its diagonal elements.

The elements of  $\mathbf{y}$  are then uncorrelated  $\langle y_i y_j \rangle = \begin{cases} 0 & , i \neq j \\ \lambda_i & , i = j \end{cases}$

## Unbiased estimate of the covariance matrix

$$\hat{C} = \frac{1}{N-1} \sum_i^N (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T$$

$$\bar{\mathbf{x}} \equiv \frac{1}{N} \sum_i^N \mathbf{x}_i$$

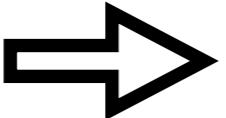


**probability of having events at**  $t_1, t_2 \dots t_n$

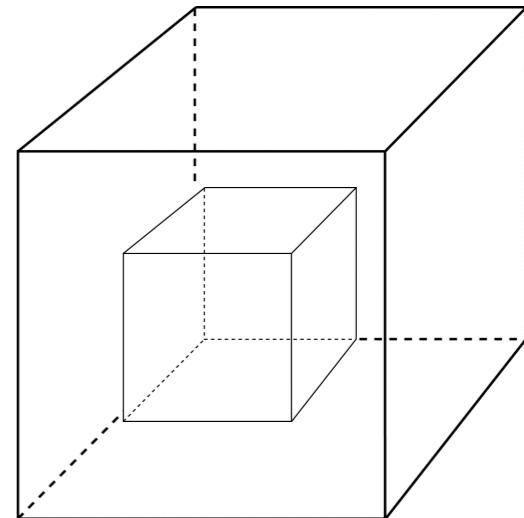
$$\begin{aligned}
p(0 < t_1 < t_2 < \dots < t_n < t) &= p(0|0, t_1) r dt_1 \Theta(t_1 < t_2) \times p(0|t_1, t_2) r dt_2 \Theta(t_2 < t_3) \\
&\quad \dots \times p(0|t_n, t) dt_n \Theta(t_n < t) \\
&= r^n e^{-rt} \prod_{i=1}^n dt_i \Theta(t_i < t_{i+1})
\end{aligned}$$

**marginalize over the individual times while keeping their order**

$$\begin{aligned}
p(n|r, t) &= \prod_i \int_0^t dt_i p(0 < t_1 < t_2 < \dots < t_n < t) \\
&= r^n e^{-rt} \int_0^t dt_n \dots \int_0^{t_3} dt_2 \int_0^{t_2} dt_1 \\
&= r^n e^{-rt} \int_0^t dt_n \dots \int_0^{t_3} dt_2 t_2 \\
&= r^n e^{-rt} \int_0^t dt_n \dots \int_0^{t_4} dt_3 \frac{t_3^2}{2} \\
&= \frac{(rt)^n}{n!} e^{-rt} \\
&= \frac{(\nu)^n}{n!} e^{-\nu}
\end{aligned}$$

**binomial**  **Poisson**

$$\begin{aligned}
 \binom{N}{n} p^n (1-p)^{N-n} &= \binom{N}{n} \left(\frac{v}{V}\right)^n \left(1 - \frac{v}{V}\right)^{N-n} \\
 &= \binom{N}{n} \left(\frac{\nu}{N}\right)^n \left(1 - \frac{\nu}{N}\right)^{N-n} \\
 &= \frac{N!}{n!(N-n)!} \left(\frac{\nu}{N}\right)^n \left(1 - \frac{\nu}{N}\right)^{N-n} \\
 &= \frac{N^n}{n!} \left(\frac{\nu}{N}\right)^n \left(1 - \frac{\nu}{N}\right)^{N-n} \\
 &\simeq \frac{\nu^n}{n!} e^{-\nu}
 \end{aligned}$$

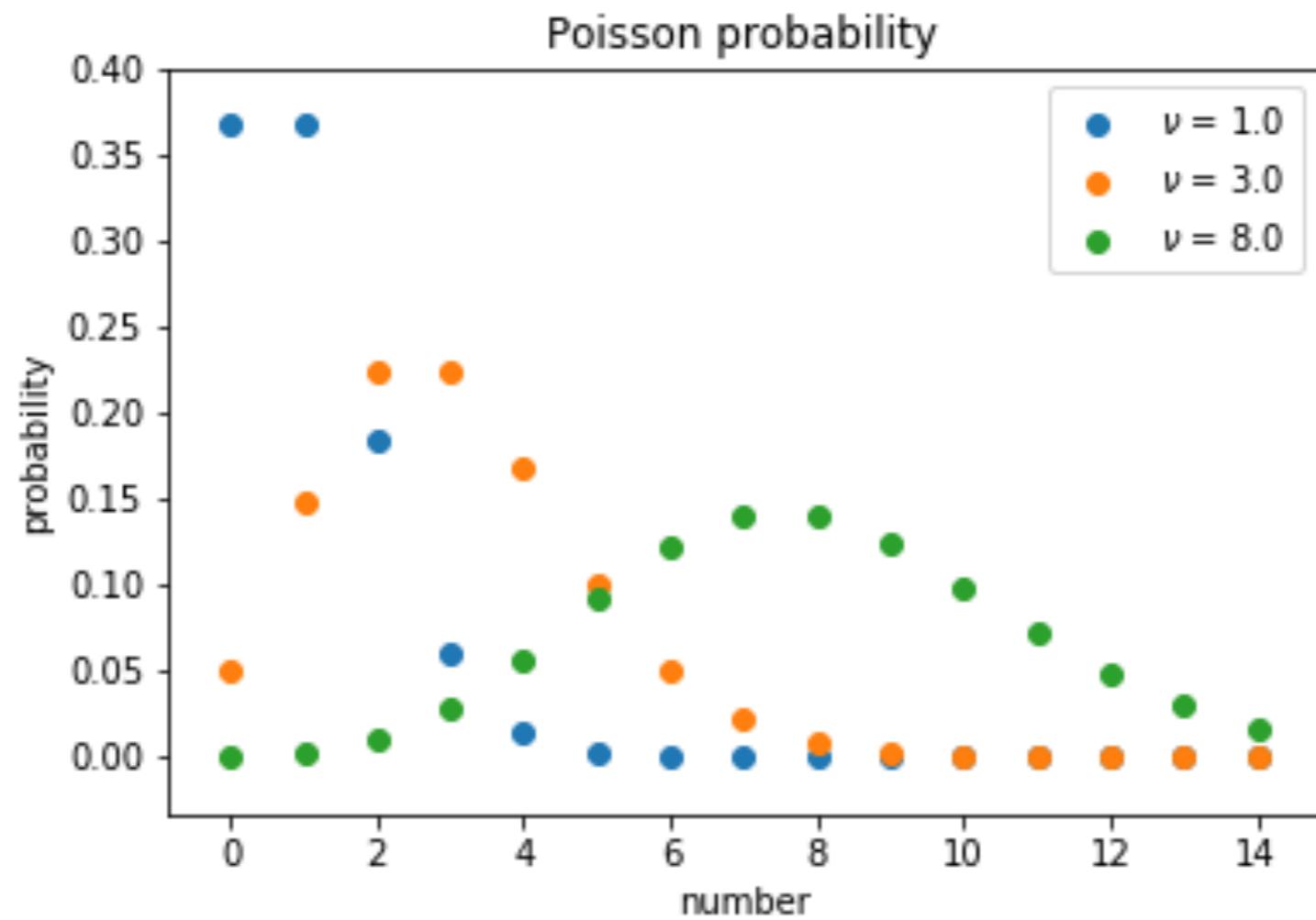


$$\eta = \frac{N}{V} \quad \text{Number density}$$

$$\nu = \eta v$$

$$\text{using } \frac{N!}{(N-n)!} \simeq N^n$$

$$\lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n = e^x$$



$$p(n|\nu) = \frac{\nu^n}{n!} e^{-\nu}$$

$$E[n] = \nu$$

$$Var[n] = \nu$$

# Gaussian or Normal Distribution

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

**Cumulative distribution :**  $F(x) = \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)$

**the error function**  $\operatorname{erf}(z) \equiv \frac{2}{\sqrt{\pi}} \int_0^z e^{-u^2} du$

# Gaussian or Normal Distribution

$$p(-\sigma \leq x - \mu \leq \sigma) = 0.683$$

$$p(-2\sigma \leq x - \mu \leq 2\sigma) = 0.954$$

$$p(-3\sigma \leq x - \mu \leq 3\sigma) = 0.997$$

$$p(-4\sigma \leq x - \mu \leq 4\sigma) = 0.999937$$

# central limit theorem

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N x_i$$

Where the  $x_i$  are independent random numbers with mean  $\mu$  and variance  $\sigma^2$  drawn from **any distribution**.

In the limit  $N \rightarrow \infty$ ,  $\bar{X}$  is normally distributed.

$$\bar{X} \xrightarrow{p} \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$$

# central limit theorem

$$S = \sum_i x_i$$

$$p(S) = \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_n \ p(S, x_1 \dots x_n) \quad (1.10.9)$$

$$= \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_n \ p(S|x_1 \dots x_n)p(x_1 \dots x_n) \quad (1.10.10)$$

$$= \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_n \ \delta(S - \sum_i x_i) \ p_1(x_1) \dots p_n(x_n) \quad (1.10.11)$$

$$= \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_n \ \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} \exp \left[ -ik(S - \sum_i x_i) \right] \ p_1(x_1) \dots p_n(x_n) \quad (1.10.12)$$

$$= \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS} \prod_i \int_{-\infty}^{\infty} dx_i e^{+ikx_i} p_i(x_i) \quad (1.10.13)$$

$$= \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS} \prod_i \tilde{p}_i(k) \quad (1.10.14)$$

$$= \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS} \tilde{p}_S(k) \quad (1.10.15)$$

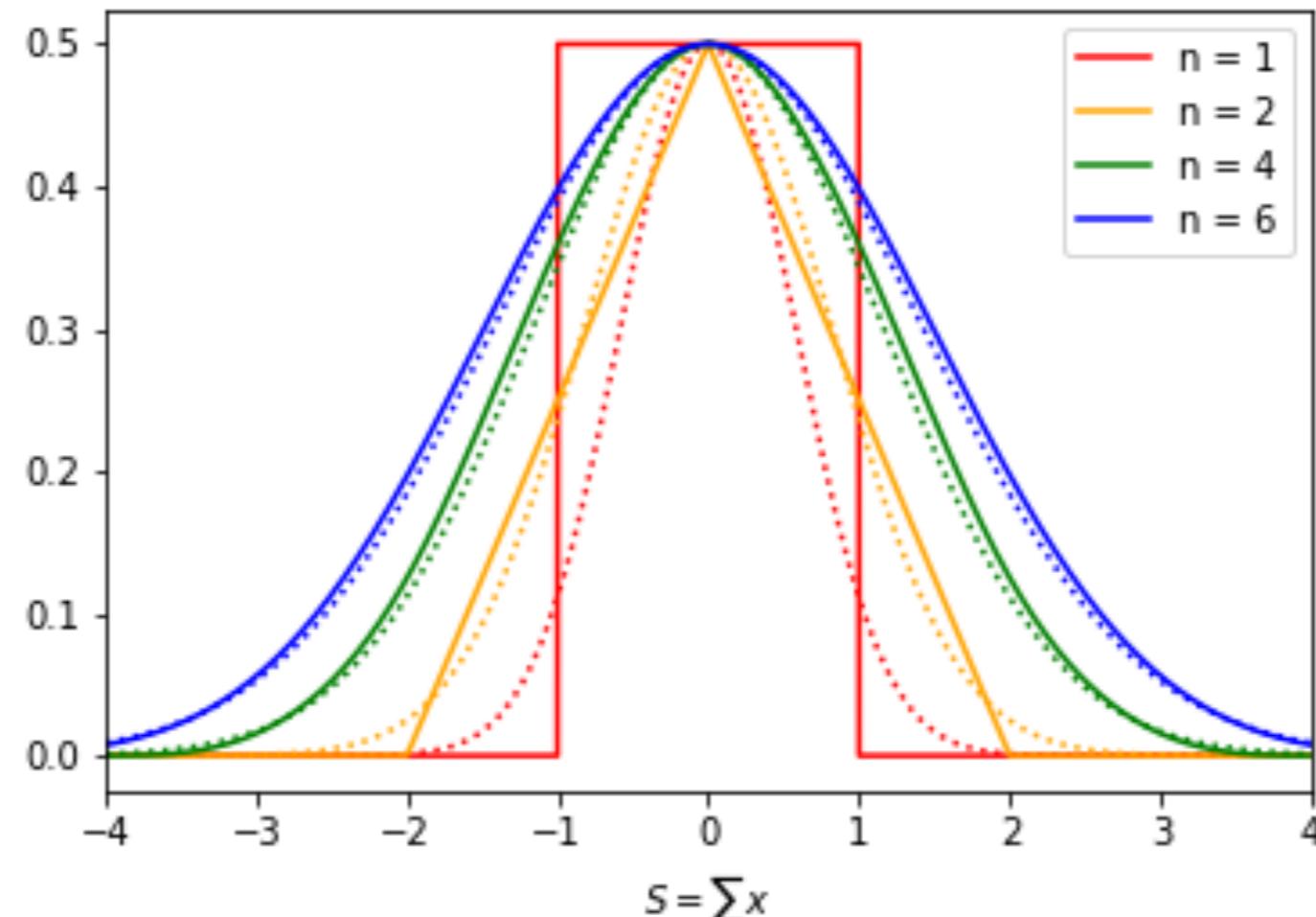
# central limit theorem

The characteristic function (Fourier transform) of uniform distribution between  $-L/2$  and  $L/2$  is:

$$\tilde{p}(k) = \frac{1}{L} \int_{-L/2}^{L/2} dx e^{+ikx} = \frac{2}{Lk} \sin\left(\frac{kL}{2}\right) = \text{sinc}\left(\frac{kL}{2}\right)$$

So the pdf for the sum of  $n$  uniformly distributed variables, each over a range  $L/n$  is

$$p_n(S) = \int_{-\infty}^{+\infty} \frac{dk}{(2\pi)} e^{-ikS} \text{sinc}^n\left(\frac{kL}{2n}\right)$$



**Figure:** Probability distribution for the sum of  $n$  random variables that are uniformly distributed between -1 and 1. The normalizations have been changed so that their maximum is 0.5 in all cases. The dotted curves are for Gaussians with the same variance. You can see that the distribution converges to Gaussian remarkably quickly even for a very non Gaussian initial distribution.



# Connection between Gaussian and Poisson Distributions

Poisson distribution

$$p(n|\nu) = \frac{(\nu)^n}{n!} e^{-\nu}$$

substitution  $n = \nu(1 + \delta)$ ,  $\delta = (n - \nu)/\nu$ .

$$n! \sim \sqrt{2\pi n} e^{-n} n^n \quad n \gg 1$$

$$\begin{aligned} p(n) &= \frac{\nu^{\nu(1+\delta)} e^{-\nu}}{\sqrt{2\pi} e^{-\nu(1+\delta)} [\nu(1 + \delta)]^{\nu(1+\delta)+1/2}} \\ &= \frac{e^{\nu\delta}(1 + \delta)^{-\nu(1+\delta)-1/2}}{\sqrt{2\pi\nu}} \end{aligned}$$

Let's look at the lowest order terms of the log of the numerator

$$\ln \left[ (1 + \delta)^{-\nu(1+\delta)-1/2} \right] = -(\nu(1 + \delta) + 1/2) \ln(1 + \delta) \quad (36)$$

$$= -(\nu + \nu\delta + 1/2) \left( \delta - \frac{\delta^2}{2} + \dots \right) \quad \nu \gg 1 \quad (37)$$

$$\simeq -(\nu + \nu\delta) \left( \delta - \frac{\delta^2}{2} + \dots \right) \quad (38)$$

$$\simeq -\nu\delta - \frac{\nu\delta^2}{2} + \dots \quad (39)$$

Putting this back into the above

$$p(\delta) = p(n) \quad (40)$$

$$\simeq \sqrt{\frac{\nu}{2\pi}} e^{-\frac{\nu\delta^2}{2}}. \quad (41)$$

So if  $\nu$  is large the excursion from the mean,  $\delta$ , is Gaussian distributed with a variance of  $1/\nu$ .

# Multivariate Gaussian

$$p(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C}) = \frac{1}{(2\pi)^{n/2} \sqrt{|\mathbf{C}|}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] = \mathcal{G}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C})$$

Theorem

$$E[x_i] = \mu_i \quad \text{or} \quad E[\mathbf{x}] = \boldsymbol{\mu}$$

Theorem

$$\begin{aligned} \text{Cov}[x_i x_j] &= E[(x_i - \mu_i)(x_j - \mu_j)] = C_{ij} \\ \text{or } \text{Cov}[\mathbf{x}\mathbf{x}] &= E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \mathbf{C} \end{aligned}$$

Theorem

If  $\mathbf{x} \sim \mathcal{N}(0, \mathbf{C}_1)$  and  $\mathbf{x}' \sim \mathcal{N}(0, \mathbf{C}_2)$  and their sum is  $\mathbf{s} = \mathbf{x} + \mathbf{x}'$  then  $\mathbf{s} \sim \mathcal{N}(0, \mathbf{C}_1 + \mathbf{C}_2)$ .

# Multivariate Gaussian

$$\mathbf{x} = \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_z \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} \mathbf{C}_{yy} & \mathbf{C}_{zy} \\ \mathbf{C}_{zu}^T & \mathbf{C}_{zz} \end{bmatrix}$$

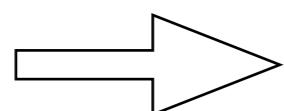
- conditional probability

$$p(\mathbf{y}|\mathbf{z}) = \mathcal{G}(\mathbf{y} | \boldsymbol{\mu}'_y, \boldsymbol{\Sigma}_{yy}) \quad \begin{cases} \boldsymbol{\mu}'_y = \boldsymbol{\mu}_y + \mathbf{C}_{zy} \mathbf{C}_{zz}^{-1} (\mathbf{z} - \boldsymbol{\mu}_z) \\ \boldsymbol{\Sigma}_{yy} = \mathbf{C}_{yy} - \mathbf{C}_{zy} \mathbf{C}_{zz}^{-1} \mathbf{C}_{zy}^T \end{cases}$$

- marginal probability

integrate over the parameters  $\mathbf{z}$  we get the marginal distribution

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



$$p(\mathbf{y}) = \mathcal{G}(\mathbf{y} | \boldsymbol{\mu}_y, \mathbf{C}_{yy})$$

# Multivariate Gaussian

**Any linear (affine) transformation of normally distributed variables will be normally distributed.**

If  $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$  ,  $\mathbf{c}$  and  $\mathbf{A}$  are constant, and

$$\mathbf{y} = \mathbf{Ax} + \mathbf{c} \quad \text{i.e. } \left( y_j = \sum_i a_{ji}x_i + c_j \right)$$

then  $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_y, \mathbf{C}_y)$

$$z = [\chi^2] = \sum_i^N \frac{(x_i - \mu_i)^2}{\sigma_i^2}. \quad (61)$$

The Gaussian distribution is

$$p(x_1, x_2, \dots, x_N) dx_1 \dots dx_N = \frac{1}{(2\pi)^{N/2} \prod_i \sigma_i} e^{-\frac{1}{2} \sum_i^N \frac{(x_i - \mu_i)^2}{\sigma_i^2}} dx_1 \dots dx_N \quad (62)$$

$$= \frac{1}{(2\pi)^{N/2} \prod_i \sigma_i} e^{-\frac{1}{2} z} dx_1 \dots dx_N \quad (63)$$

$$= \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2} \sum_{i=1}^N x_i'^2} dx'_1 \dots dx'_N \quad x' = \frac{x - \mu}{\sigma} \quad (64)$$

$$= \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2} z} dx'_1 \dots dx'_N \quad (65)$$

$$dx'_1 \dots dx'_N = r^{n-1} dr d\theta_1 d\theta_3 \dots = \frac{1}{2} z^{n/2-1} dz d^n \Omega \quad (66)$$

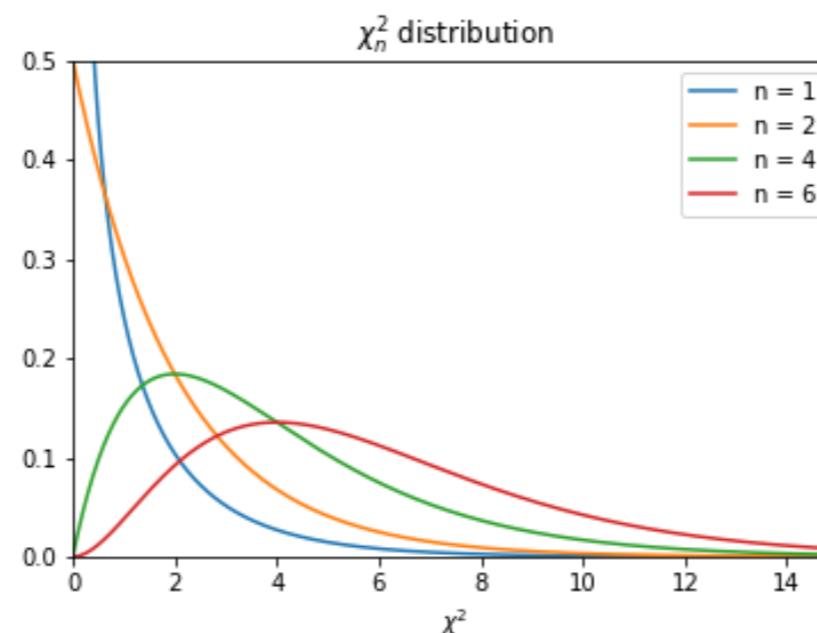
# $\chi^2$ distribution

$$p(z = \chi^2 | n) = \begin{cases} \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} z^{\frac{n}{2}-1} e^{-\frac{z}{2}} & z \geq 0 \\ 0 & z < 0 \end{cases}$$

**gamma function**  $\Gamma(x) \equiv \int_0^\infty dt e^{-t} t^{x-1}$ ,  $\Gamma(n+1) = n!$

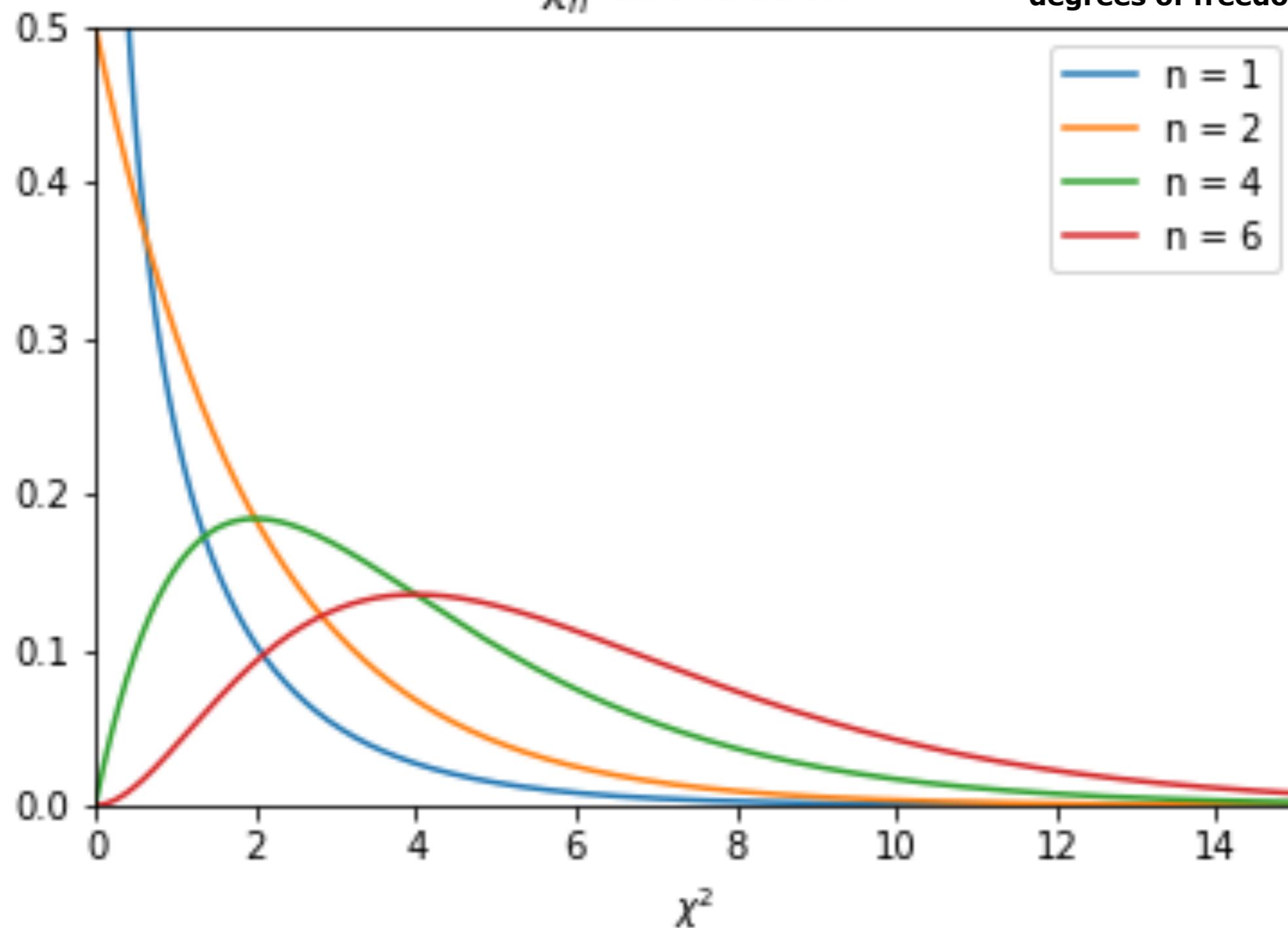
$$E[x] = n \quad \text{Var}[x] = 2n \quad \max_z p(z|n) = \max(n-2, 0)$$

For this reason the value of  $\chi_n^2/n$  is often given and compared to 1.



# $\chi_n^2$ distribution

degrees of freedom



# Law of Large Numbers

$$\lim_{n \rightarrow \infty} \left[ \frac{1}{n} \sum_{i=1}^n f(x_i) \right] = E [f(x)]$$

**Statistic** - any function of numbers drawn from a random distribution.

**Estimators** - a statistic,  $\hat{\theta}(x_1, \dots, x_n)$ , that is meant to estimate the value of some quantity of interest,  $\theta$ .

**Statistic** - any function of numbers drawn from a random distribution.

**Estimators** - a statistic,  $\hat{\theta}(x_1, \dots, x_n)$ , that is meant to estimate the value of some quantity of interest,  $\theta$ .

Questions to be asked about an estimator :

Is the estimator accurate?

$$E[\hat{\theta}(x_1, \dots, x_n)] = \theta \quad \text{unbiased estimator}$$

$$E[\hat{\theta}(x_1, \dots, x_n)] \neq \theta \quad \text{biased estimator}$$

**Statistic** - any function of numbers drawn from a random distribution.

**Estimators** - a statistic,  $\hat{\theta}(x_1, \dots, x_n)$ , that is meant to estimate the value of some quantity of interest,  $\theta$ .

Questions to be asked about an estimator :

Is the estimator precise?

$$E \left[ (\hat{\theta}(x_1, \dots, x_n) - \theta)^2 \right] \quad \text{variance of the estimator}$$

or  $E \left[ (\hat{\theta} - E[\hat{\theta}])^2 \right]$  if it is biased.

$$\bar{x}_N \equiv \frac{1}{N} \sum_{i=0}^N x_i$$

$$\langle \bar{x}_N \rangle = \frac{1}{N} \sum_{i=0}^N \langle x_i \rangle$$

$$= \frac{1}{N} \sum_{i=0}^N \mu$$

$$= \mu \quad \text{unbiased}$$

## Is the estimator precise?

$$\begin{aligned}
Var[\bar{x}_N] &= \langle [\bar{x}_N - \mu]^2 \rangle \\
&= \langle [\bar{x}_N]^2 \rangle - 2\mu \langle \bar{x}_N \rangle + \mu^2 \\
&= \langle [\bar{x}_N]^2 \rangle - \mu^2 \\
&= \left\langle \left[ \frac{1}{N} \sum_{i=0}^N x_i \right]^2 \right\rangle - \mu^2 \\
&= \frac{1}{N^2} \sum_{i=0}^N \sum_{j=0}^N \langle x_i x_j \rangle - \mu^2 \\
&= \frac{1}{N^2} \left[ \sum_{i=0}^N \langle x_i^2 \rangle + \sum_{i \neq j} \langle x_i x_j \rangle \right] - \mu^2 \\
&= \frac{1}{N^2} \left[ \sum_{i=0}^N (\sigma^2 + \mu^2) + \sum_{i \neq j} \langle x_i \rangle \langle x_j \rangle \right] - \mu^2 \\
&= \frac{1}{N^2} [N(\sigma^2 + \mu^2) + N(N-1)\mu^2] - \mu^2 \\
&= \frac{\sigma^2}{N}
\end{aligned}$$

Consider the estimator

$$\hat{\theta} = \sum_i w_i x_i \quad , \quad \sum_i w_i = 1 \quad (80)$$

The variance of the estimator will be

$$\begin{aligned}\sigma_{\theta}^2 &= \langle \theta^2 \rangle - \mu^2 \\ &= \left\langle \left[ \sum_i w_i x_i \right]^2 \right\rangle - \mu^2 \\ &= \sum_{ij} w_i w_j \langle x_i x_j \rangle - \mu^2 \\ &= \sum_i w_i^2 \langle x_i^2 \rangle + \sum_{i \neq j} w_i w_j \langle x_i \rangle \langle x_j \rangle - \mu^2 \\ &= \sum_i w_i^2 [\sigma_i^2 + \mu^2] + \mu^2 \sum_{i \neq j} w_i w_j - \mu^2\end{aligned}$$

$$F(\mathbf{w}) = \sigma_\theta^2(\mathbf{w}) + \lambda \left(1 - \sum_i w_i\right)$$

that is

$$\frac{\partial F}{\partial w_k} = \frac{\partial \sigma_\theta^2}{\partial w_k} - \lambda = 0 \quad (87)$$

$$\begin{aligned}
 \frac{\partial \sigma_\theta^2}{\partial w_k} &= 2w_k [\sigma_k^2 + \mu^2] + 2\mu^2 \sum_{i \neq k} w_i \\
 &= 2w_k [\sigma_k^2 + \mu^2] + 2\mu^2 \left[ \sum_{i=0}^N w_i - w_k \right] \\
 &= 2w_k [\sigma_k^2 + \mu^2] + 2\mu^2 [1 - w_k] \quad \text{use constraint } \sum_i w_i = 1 \\
 &= 2w_k \sigma_k^2 + 2\mu^2
 \end{aligned}$$

putting this into (87) gives

$$w_k = \frac{\lambda - 2\mu}{2\sigma_k^2}$$

Plugging this into the constraint

$$\lambda = 2\mu + 2 \left[ \sum_k \frac{1}{\sigma_k^2} \right]^{-1}$$

so

$$w_k = \left[ \sum_i \frac{1}{\sigma_i^2} \right]^{-1} \frac{1}{\sigma_k^2}$$

So the estimator (80), the one with the minimum variance, is

$$\hat{\theta} = \frac{1}{\left[ \sum_i \frac{1}{\sigma_i^2} \right]} \sum_i \frac{x_i}{\sigma_i^2}.$$

# estimating the variance

$$S_N^2 = \frac{1}{N} \sum_i (x_i - \mu)^2$$

$$S_N^2 = \frac{1}{N-1} \sum_i (x_i - \bar{x}_N)^2.$$

Why is there an  $N - 1$  instead of an  $N$  in the denominator?

$$\begin{aligned}
\langle S_N^2 \rangle &= \frac{1}{N-1} \sum_i \left\langle (x_i - \bar{x}_N)^2 \right\rangle \\
&= \frac{1}{N-1} \left[ \sum_i \langle x_i^2 \rangle - 2 \left\langle \sum_i x_i \bar{x}_N \right\rangle + \sum_i \langle (\bar{x}_N)^2 \rangle \right] \\
&= \frac{1}{N-1} \left[ \sum_i (\sigma^2 + \mu^2) - 2N \langle (\bar{x}_N)^2 \rangle + N \langle (\bar{x}_N)^2 \rangle \right] \\
&= \frac{1}{N-1} \left[ \sum_i (\sigma^2 + \mu^2) - N \langle (\bar{x}_N)^2 \rangle \right] \\
&= \frac{1}{N-1} \left[ N(\sigma^2 + \mu^2) - N \left( \frac{\sigma^2}{N} + \mu^2 \right) \right] \\
&= \sigma^2
\end{aligned}$$

So this estimator is unbiased. Note that this does not require that the  $x$ 's be normally distributed.

If there were an  $N$  in the denominator of (102) then  $\langle s_N^2 \rangle = (N-1)\sigma/N$  which means it would be **biased**, but since the bias gets smaller as  $N$  increases it would be a simple example of an **asymptotically unbiased estimator**.

**Theorem**

If  $x_i \sim \mathcal{N}(\mu, \sigma)$  and  $S_N$  is given by (102) then  $z = \frac{(N-1)S_N^2}{\sigma^2}$  is  $\chi_{N-1}^2$  distributed.

**half proof:**

$$\begin{aligned}
 \frac{(N-1)S_N^2}{\sigma^2} &= \frac{1}{\sigma^2} \sum_i (x_i - \bar{x})^2 \\
 &= \frac{1}{\sigma^2} \sum_i [(x_i - \mu) - (\bar{x} - \mu)]^2 \\
 &= \frac{1}{\sigma^2} \sum_i [(x_i - \mu)^2 - 2(x_i - \mu)(\bar{x} - \mu) + (\bar{x} - \mu)^2] \\
 &= \frac{1}{\sigma^2} \sum_i [(x_i - \mu)^2] - 2N(\bar{x} - \mu)(\bar{x} - \mu) + N(\bar{x} - \mu)^2 \\
 &= \sum_i \frac{(x_i - \mu)^2}{\sigma^2} - \frac{N(\bar{x} - \mu)^2}{\sigma^2}
 \end{aligned}$$

From what we know about the  $\chi^2$ , distribution, this means that our statistic  $S_N^2$  has the following properties from

$$\left\langle \frac{(N-1)}{\sigma^2} S_N^2 \right\rangle = N-1 \Rightarrow \quad \left\langle S_N^2 \right\rangle = \sigma^2$$

$$\begin{aligned} \text{Var} \left[ \frac{(N-1)}{\sigma^2} S_N^2 \right] &= \left\langle \left( \frac{(N-1)}{\sigma^2} S_N^2 \right)^2 \right\rangle - \left\langle \frac{(N-1)}{\sigma^2} S_N^2 \right\rangle^2 \\ &= 2(N-1) \\ \Rightarrow \text{Var} [S_N^2] &= \frac{2\sigma^4}{(N-1)} \end{aligned}$$

So the standard deviation of our estimated variance goes down like  $\sim 1/\sqrt{N}$ . We can also find the probability that  $S_N^2$  will be within some range using the cumulative distribution for a  $\chi^2$  distribution

$$P \left( \frac{\sigma^2}{(N-1)} z_1 < S_N^2 < \frac{\sigma^2}{(N-1)} z_2 \right) = F_{\chi_{N-1}^2}(z_2) - F_{\chi_{N-1}^2}(z_1)$$

# Estimating the Mean when the Variance is Unknown

We can estimate the mean using  $\bar{x} = \frac{1}{N} \sum_i x_i$

For normally distributed data, the sample mean is normally distributed with an variance of  $\frac{\sigma^2}{\sqrt{n}}$

We often don't know  $\sigma$  but we can estimate it using the statistic

$$S_n = \frac{1}{N-1} \sum_i (x_i - \bar{x})^2$$

But since  $\bar{x}$  is calculated from the same data the estimate of the mean will not be independent of the estimate of the variance.

The traditional solution to this is the "student-t test".

## Estimating the mean when the variance is unknown

### Theorem

If  $x_i \sim \mathcal{N}(\mu, \sigma)$  then

$$z = (\bar{x} - \mu) \sqrt{\frac{n}{S_n^2}} \quad (109)$$

is student-t distributed with  $n - 1$  degrees of freedom.

From what we know of the t distribution

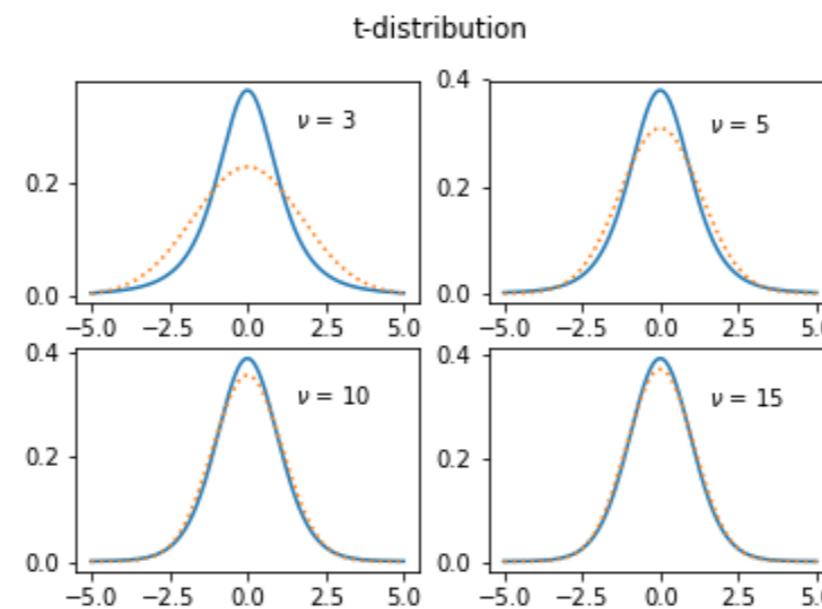
$$\text{Var}[z] = \frac{n-1}{n-3}$$

$$\text{Var}[\bar{x}] \sim \left[ \frac{n-1}{n-3} \right] \frac{S_n^2}{n}$$

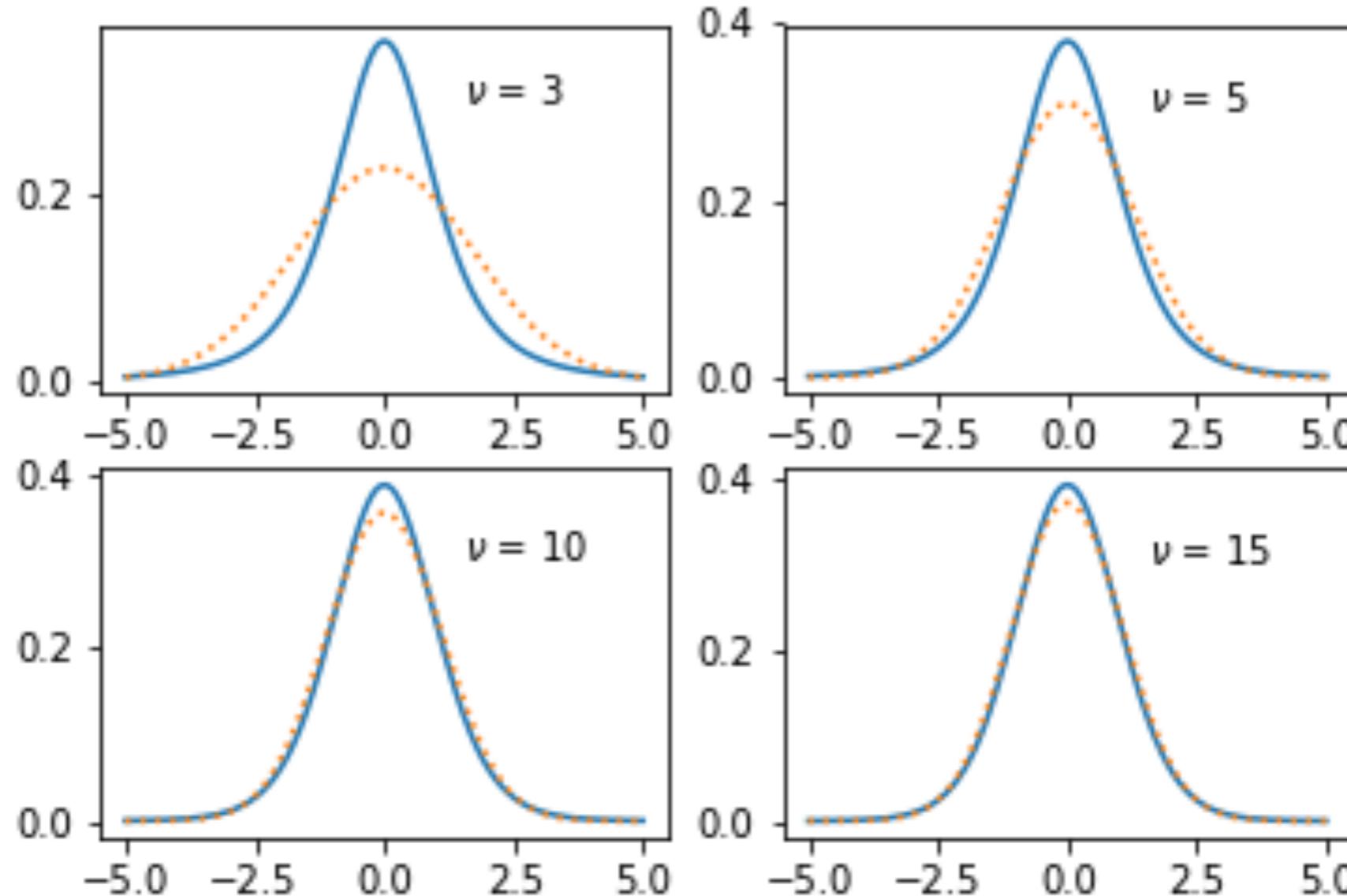
# student's t distribution

$$p_t(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left[1 + \frac{x^2}{\nu}\right]^{-\frac{\nu+1}{2}}$$

$$E[t] = 0 \quad \text{Var}[t] = \frac{\nu}{\nu - 2} \quad \text{for } \nu > 2$$



### t-distribution



## Unbiased estimate of the covariance matrix

$$\hat{C} = \frac{1}{N-1} \sum_i^N (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T$$

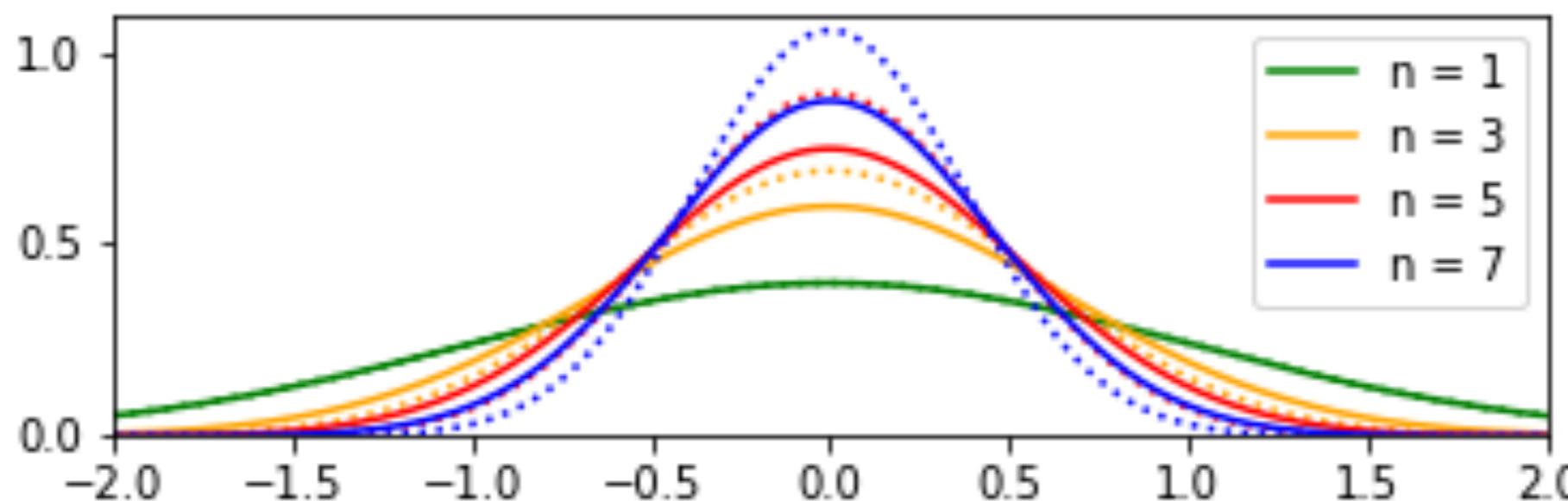
$$\bar{\mathbf{x}} \equiv \frac{1}{N} \sum_i^N \mathbf{x}_i$$

## **Problem :**

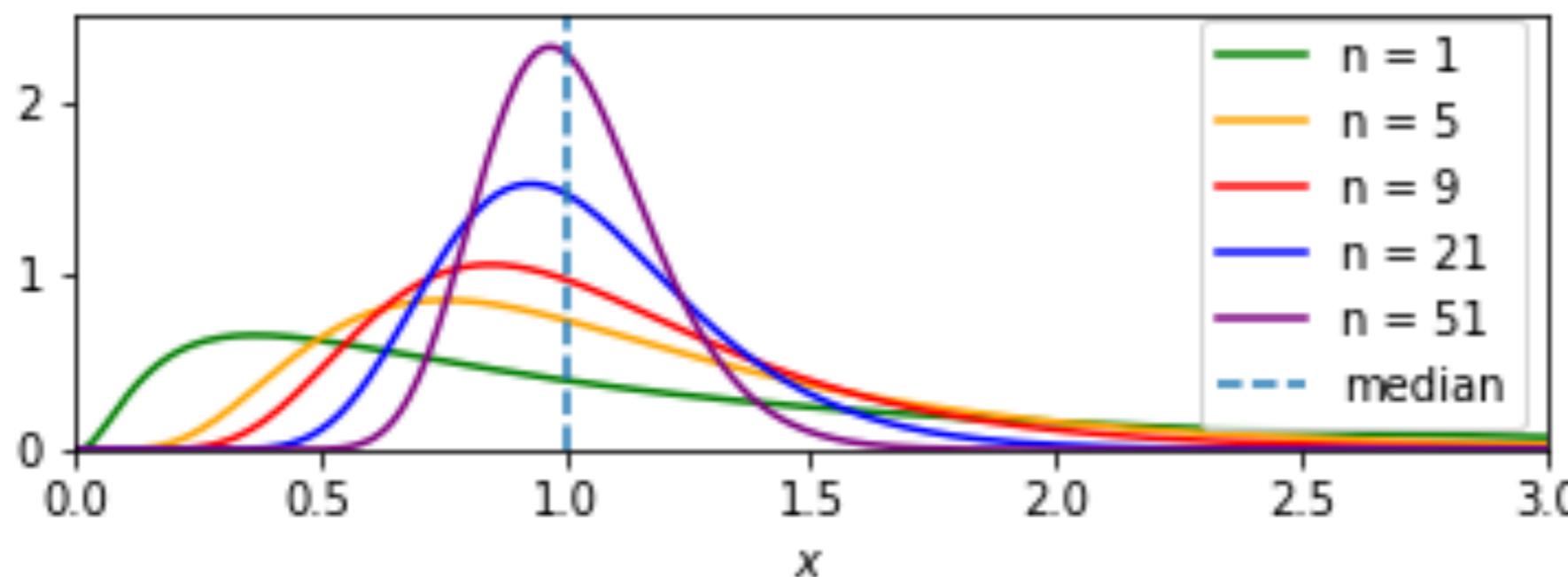
We can find the distribution of a statistic (or estimator) in terms of the parameters of the underlying distribution ( $\sigma^2$ ,  $\mu$ , ....), but we don't know what these are!

We can estimate them, but we can't say how good our estimate is.

## Distribution of the Median



Sampled from  
normal distribution



Sampled from log  
normal distribution



# Bayesian Inference

$$\begin{aligned} P(M_i | \mathcal{D}, I) &= \frac{P(\mathcal{D}|M_i, I)P(M_i|I)}{P(\mathcal{D}|I)} \\ &= \frac{P(\mathcal{D}|M_i, I)P(M_i|I)}{\sum_i P(\mathcal{D}|M_i, I)P(M_i|I)} \end{aligned}$$

$M_i$  - the model

$\mathcal{D}_i$  - the data

$$P(\mathcal{D}|I) = \sum_i P(\mathcal{D}, M_i|I) = \sum_i P(\mathcal{D}|M_i, I)P(M_i|I)$$

- $P(M_i|\mathcal{D})$  is called the **posterior probability** for model  $M_i$  given the data. This is the goal of Bayesian inference although one often summarizes this result by finding the average, mode, covariance or confidence regions.
- $P(\mathcal{D}|M_i)$  is called the **likelihood**. It is the probability of getting the observed data given the model  $M_i$ . It is often denoted  $\mathcal{L}(\mathcal{D}|M_i)$ . This is the same probability as is used in frequentist methods. Often this is a Gaussian, but not always. It includes the model that relates the parameters to the data and the description of the noise.
- $P(M_i)$  is called the **prior**. It is the probability of the model prior to the data  $\mathcal{D}$  being considered. This might take into account some previous experiment with data  $\mathcal{D}'$  in which case it would be the posterior of that experiment  $P(M_i|\mathcal{D}')$ . It might also take into account that some models, or range of parameters, are not possible in which case  $P(M_i) = 0$  for some  $i$ . For example, the mass of a planet cannot be negative or  $\Omega_{\text{matter}}$  cannot be greater than one. The prior is often denoted by  $\pi(M_i)$  in the literature.
- $P(\mathcal{D}) = \sum_i P(\mathcal{D}|M_i, I)P(M_i|I)$  is called the **evidence**. Note that the evidence is not a function of  $M_i$  although it is implicitly dependent on the set of all models considered. Since the data does not change the evidence will be a constant for a fixed set of models. We will sometimes denote the evidence as  $\mathcal{E}(\mathcal{D})$ .

$$p(M|D_1) = \frac{p(M)p(D_1|M)}{p(D_1)} \quad p(M|D_2) = \frac{p(M)p(D_2|M)}{p(D_2)}$$

Now let's look at the posterior for both data sets,

$$\begin{aligned} p(M|D_1, D_2) &= \frac{p(M)p(D_1, D_2|M)}{p(D_1, D_2)} \\ &= \frac{p(M)p(D_1|M)p(D_2|D_1, M)}{p(D_1, D_2)} \quad \text{product rule} \end{aligned}$$

$$p(\mathcal{D}_1, \mathcal{D}_2) = p(\mathcal{D}_1)p(\mathcal{D}_2)$$

$$p(\mathcal{D}_2|\mathcal{D}_1, M) = p(\mathcal{D}_2|M)$$

$$\begin{aligned} p(M|\mathcal{D}_1, \mathcal{D}_2) &= \frac{p(M)p(\mathcal{D}_1|M)p(\mathcal{D}_2|M)}{p(\mathcal{D}_1)p(\mathcal{D}_2)} \\ &= \frac{p(M)p(\mathcal{D}_1|M)}{p(\mathcal{D}_1)} \frac{p(\mathcal{D}_2|M)}{p(\mathcal{D}_2)} \\ &= p(M|\mathcal{D}_1) \frac{p(\mathcal{D}_2|M)}{p(\mathcal{D}_2)} \end{aligned}$$

Inference from one data set can be used as the prior for another data set and vice versa.

# Bayesian inference or Bayesian Parameter Estimation

$$P(\theta_1, \theta_2, \dots | D) = \frac{\mathcal{L}(D | \theta_1, \theta_2, \dots) p(\theta_1, \theta_2, \dots)}{\left[ \int d^n \theta \mathcal{L}(D | \theta_1, \theta_2, \dots) p(\theta_1, \theta_2, \dots) \right]}$$

**evidence**

posterior

likelihood

prior

The diagram illustrates the Bayesian posterior formula. It features a central equation:  $P(\theta_1, \theta_2, \dots | D) = \frac{\mathcal{L}(D | \theta_1, \theta_2, \dots) p(\theta_1, \theta_2, \dots)}{\left[ \int d^n \theta \mathcal{L}(D | \theta_1, \theta_2, \dots) p(\theta_1, \theta_2, \dots) \right]}$ . Above the equation, three labels are positioned with arrows: "posterior" points to the leftmost term  $P(\theta_1, \theta_2, \dots | D)$ , "likelihood" points to the term  $\mathcal{L}(D | \theta_1, \theta_2, \dots)$ , and "prior" points to the term  $p(\theta_1, \theta_2, \dots)$ . Below the equation, a horizontal brace spans the entire denominator  $\left[ \int d^n \theta \mathcal{L}(D | \theta_1, \theta_2, \dots) p(\theta_1, \theta_2, \dots) \right]$ , with the word "evidence" centered below it.

# radioactive decay example

Likelihood:

$$p(n|r) = \frac{(r\delta t)^n}{n!} e^{-r\delta t} = \mathcal{L}(n|r) \quad \text{Poisson}$$

Uniform prior:

$$p(r) = \pi(r) = \frac{\Theta(0 < r < r_{max})}{r_{max}}$$

Evidence:

$$\begin{aligned} \mathcal{E}(n) &= \int_{-\infty}^{\infty} dr \ p(n|r)p(r) = \frac{1}{r_{max}} \int_0^{r_{max}} dr \ \frac{(r\delta t)^n}{n!} e^{-r\delta t} \\ &= \frac{\delta t^{-1}}{n!r_{max}} \int_0^{\delta t r_{max}} dx \ x^n e^{-x} \quad x = r\delta t \\ &\simeq \frac{\delta t^{-1}}{n!r_{max}} \int_0^{\infty} dx \ x^n e^{-x} \quad r_{max} \gg 1/\delta t \\ &= \frac{\delta t^{-1}}{n!r_{max}} \Gamma(n+1) \\ &= \frac{1}{\delta t r_{max}} \end{aligned}$$

because  $\Gamma(n+1) = n!$

**Posterior** for the rate is

$$p(r|n) = \frac{\delta t}{n!} (\delta t r)^n e^{-r\delta t}$$

**Average:**

$$\begin{aligned}\langle r \rangle &= \int_0^\infty dr \ r p(r|n) = \frac{\delta t}{n!} \int_0^\infty dr \ r (\delta t r)^n e^{-r\delta t} = \frac{1}{\delta t n!} \int_0^\infty dx \ x^{n+1} e^{-x} \\ &= \frac{(n+1)!}{\delta t n!} = \frac{(n+1)}{\delta t}\end{aligned}$$

**Variance:**

$$Var[r] = \frac{(n+1)}{\delta t^2}$$

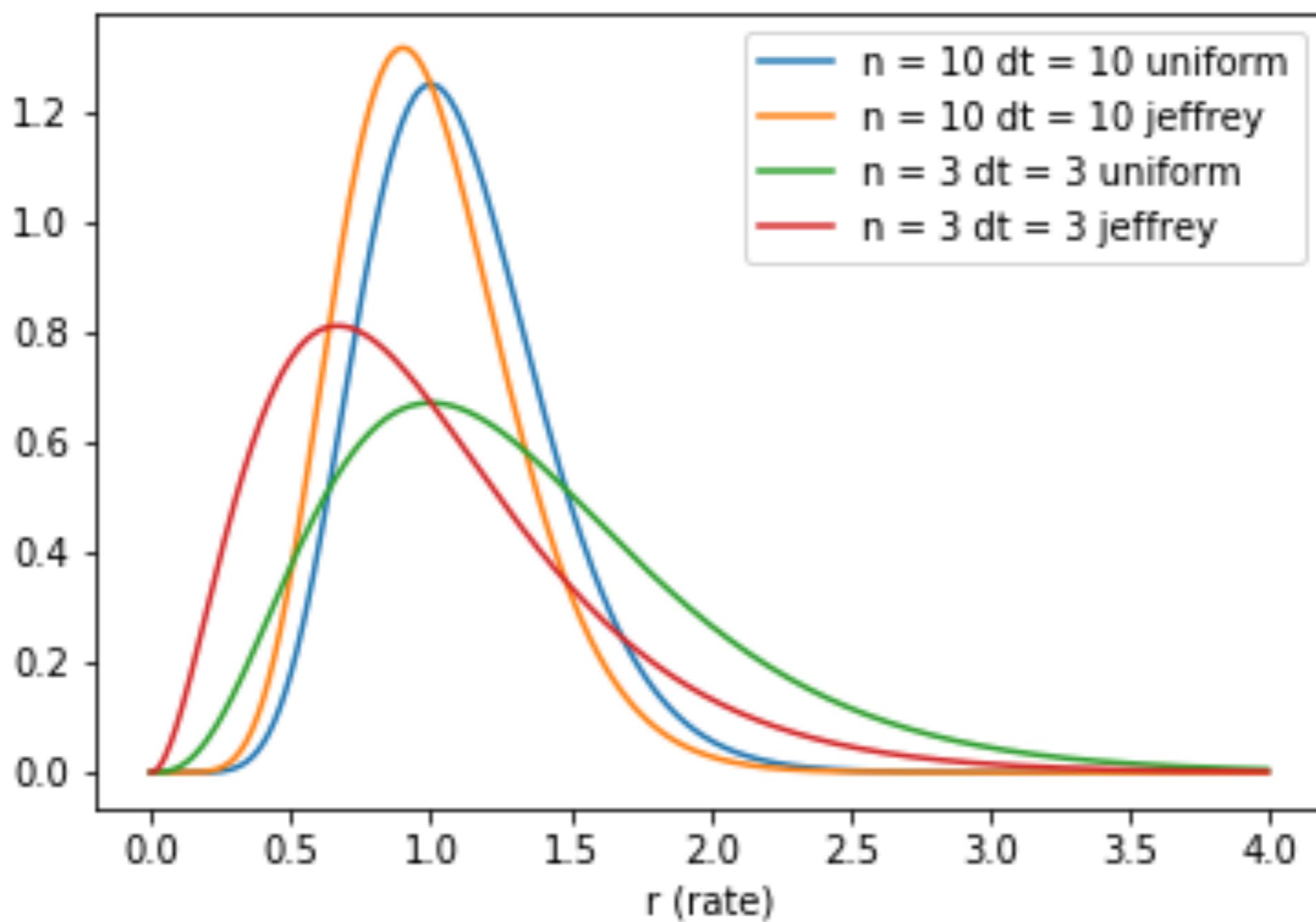
**Mode**

$$\begin{aligned}\frac{\partial}{\partial r} \ln p(r|n) &= \frac{\partial}{\partial r} ([n \ln(r\delta t) - r\delta t - \ln(\delta t/n!)]) \\ &= \frac{n}{r} - \delta t\end{aligned}$$

so the most likely value is what we might have expected,

$$r_{mode} = \frac{n}{\delta t}.$$

**maximum posterior estimate** (MPE or MAP) for  $r$  which in the case of a uniform prior is also the **maximum likelihood estimator** (MLE).



# example: estimating the mean

Additive noise model

$$d_i = \theta + n_i,$$

the data is some fixed value plus a noise component. The likelihood will be

$$\mathcal{L}(d|\theta) = \mathcal{G}(d|\theta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(d-\theta)^2}{2\sigma^2}}$$

Uniform prior

$$\begin{aligned} p(\theta) &= \begin{cases} \frac{1}{\theta_{\max} - \theta_{\min}} & \theta_{\min} < \theta < \theta_{\max} \\ 0 & \text{otherwise} \end{cases} \\ &= \mathcal{C}\Theta(\theta_{\min} < \theta < \theta_{\max}) \quad \mathcal{C} \equiv \frac{1}{\theta_{\max} - \theta_{\min}} \end{aligned}$$

So in that case the posterior is equal to the likelihood,  $\mathcal{G}(d|\theta, \sigma)$  which obviously has a mode at  $\theta = d$  and the average is  $\langle \theta \rangle = d$ .

For  $N$  data points

$$\begin{aligned} \mathcal{L}(\mathbf{d}|\theta) &= \mathcal{G}(d_1|\theta, \sigma) \times \mathcal{G}(d_2|\theta, \sigma) \times \dots \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2} \sum_i \frac{(d_i - \theta)^2}{\sigma^2}\right) \end{aligned}$$

$$\begin{aligned}
 \mathcal{L}(\mathbf{d}|\theta) &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_i (d_i^2 - 2d_i\theta + \theta^2)\right) \\
 &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \left[ \sum_i d_i^2 - 2 \sum_i d_i\theta + n\theta^2 \right]\right) \\
 &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \left[ n\bar{d}^2 + n(\theta - \bar{d})^2 - n(\bar{d})^2 \right]\right) \\
 &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{n}{2\sigma^2} \left[ \bar{d}^2 - (\bar{d})^2 \right]\right) \exp\left(-\frac{n}{2\sigma^2}(\theta - \bar{d})^2\right)
 \end{aligned}$$

where

$$\bar{d} \equiv \frac{1}{n} \sum_i d_i \quad \bar{d}^2 \equiv \frac{1}{n} \sum_i d_i^2.$$

To find the evidence we need to integrate this over  $\theta$ .

$$\mathcal{E}(\mathbf{d}) = \frac{1}{(2\pi)^{(n-1)/2}\sigma^{n-1}\sqrt{n}} \exp\left(-\frac{n}{2\sigma^2} \left[ \bar{d}^2 - (\bar{d})^2 \right]\right)$$

The posterior is

$$P(\theta|\mathbf{d}) = \sqrt{\frac{n}{2\pi\sigma^2}} \exp\left(-\frac{n}{2\sigma^2}(\theta - \bar{d})^2\right) = \mathcal{G}(\theta | \bar{d}, \sigma^2/n).$$

The posterior is

$$P(\theta|\mathbf{d}) = \sqrt{\frac{n}{2\pi\sigma^2}} \exp\left(-\frac{n}{2\sigma^2}(\theta - \bar{d})^2\right) = \mathcal{G}(\theta | \bar{d}, \sigma^2/n).$$

$$\bar{d} \equiv \frac{1}{n} \sum_i d_i \quad \bar{d^2} \equiv \frac{1}{n} \sum_i d_i^2.$$

$\bar{d}$  is a sufficient statistic for the parameter  $\theta$

$t(\mathbf{d})$  is a **sufficient statistic** for the parameter  $\theta$  if  $p(\mathbf{d}|\theta) = f(\mathbf{d})p(t(\mathbf{d})|\theta)$

which implies that the posterior can be written  $p(\theta|\mathbf{d}) = p(\theta|t(\mathbf{d}))$

# example: estimating the mean and variance

$$d_i = \theta + x_i + n_i$$

Likelihood

$$\begin{aligned}\mathcal{L}(\mathbf{d} | \theta, \sigma_n^2, \sigma_a^2) &= \int_{-\infty}^{\infty} d^n x \ P(\mathbf{d}, \mathbf{x} | \theta, \sigma_a^2) \\ &= \int_{-\infty}^{\infty} d^n x \ [ \mathcal{G}(d_1 | x_1, \sigma_n^2) \mathcal{G}(d_2 | x_2, \sigma_n^2) \dots ] [ \mathcal{G}(x_1 | \theta, \sigma_a^2) \mathcal{G}(x_2 | \theta, \sigma_a^2) \dots ] \\ &= \int_{-\infty}^{\infty} d^n x \mathcal{G}(\mathbf{d} | \mathbf{x}, \sigma_n^2) \mathcal{G}(\mathbf{x} | \theta, \sigma_a^2) \\ &= \mathcal{G}(\mathbf{d} | \theta, \sigma_n^2 + \sigma_a^2)\end{aligned}$$

Same likelihood as we got in the first example except  $\sigma^2$  is replaced with  $\sigma_n^2 + \sigma_a^2$ ,

↳ example: estimating the mean and variance

$$\mathcal{L}(\mathbf{d}|\theta, \sigma_n^2, \sigma_a^2) = \frac{1}{\sqrt{(2\pi)^{n/2}(\sigma_n^2 + \sigma_a^2)^n}} \exp\left(-\frac{n[\bar{d}^2 - (\bar{d})^2]}{2(\sigma_n^2 + \sigma_a^2)}\right) \exp\left(-\frac{n(\theta - \bar{d})^2}{2(\sigma_n^2 + \sigma_a^2)}\right)$$

To make things simpler let's make the following substitutions

$$\Delta^2 \equiv \bar{d}^2 - (\bar{d})^2$$

$$\sigma^2 \equiv \sigma_n^2 + \sigma_a^2$$

We can then use  $\sigma^2$  as a parameter instead of  $\sigma_a^2$ . The likelihood is now

$$\mathcal{L}(\mathbf{d}|\theta, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{n\Delta^2}{2\sigma^2}\right) \exp\left(-\frac{n(\theta - \bar{d})^2}{2\sigma^2}\right)$$

$\Delta^2$  and  $\bar{d}$  are sufficient for  $\sigma_a^2$  and  $\theta$

We will assume a uniform prior for both  $\theta$  and  $\sigma^2$

$$\begin{aligned} P(\theta, \sigma^2) &= \frac{\Theta(\theta_{\max} < \theta < \theta_{\min})}{(\theta_{\max} - \theta_{\min})} \frac{\Theta(0 < \sigma^2 < \sigma_{\max}^2)}{\sigma_{\max}^2} \\ &= C \Theta(\theta_{\max} < \theta < \theta_{\min}) \Theta(0 < \sigma^2 < \sigma_{\max}^2) \end{aligned}$$

evidence:

$$\begin{aligned} \mathcal{E}(\mathbf{d}) &= C \int_0^{\sigma_{\max}^2} d\sigma^2 \int_{\theta_{\min}}^{\theta_{\max}} d\theta \mathcal{L}(\mathbf{d}|\theta, \sigma^2) \\ &\simeq C \int_0^{\sigma_{\max}^2} d\sigma^2 \int_{-\infty}^{\infty} d\theta \mathcal{L}(\mathbf{d}|\theta, \sigma^2) \\ &= \frac{C}{(2\pi)^{n/2}} \int_0^{\sigma_{\max}^2} d\sigma^2 \int_{\theta_{\min}}^{\theta_{\max}} d\theta \frac{1}{\sigma^n} \exp\left(-\frac{n\Delta^2}{2\sigma^2}\right) \exp\left(-\frac{n(\theta - \bar{d})^2}{2\sigma^2}\right) \\ &= \frac{C}{(2\pi)^{(n-1)/2}} \int_0^{\sigma_{\max}^2} d\sigma^2 \frac{1}{\sigma^{n-1}} \exp\left(-\frac{n\Delta^2}{2\sigma^2}\right) \end{aligned}$$

In doing this we have taken the range of the  $\theta$  integral to go to infinity. This is justifiable if  $|\theta_{\max}| = |\theta_{\min}| \gg \sigma$ . We don't know this ahead of time, but it can be justified in retrospect once constraints on  $\sigma$  are found. This can be considered a technical flaw that we will get back to later.

Now let's make the change of variables to

$$y = \sqrt{\frac{n\Delta^2}{2\sigma^2}} \quad \text{so} \quad d\sigma^2 = \frac{n\Delta^2}{y^3} dy$$

$$\begin{aligned} \mathcal{E}(\mathbf{d}) &= \frac{2\mathcal{C}}{(2\pi)^{(n-1)/2}} \left( \frac{n\Delta^2}{2} \right)^{\frac{3-n}{2}} \int_{\sqrt{\frac{n\Delta^2}{2\sigma_{\max}^2}}}^{\infty} dy \ y^{n-4} e^{-y^2} \\ &\simeq \frac{2\mathcal{C}}{(2\pi)^{(n-1)/2}} \left( \frac{n\Delta^2}{2} \right)^{\frac{3-n}{2}} \int_0^{\infty} dy \ y^{n-4} e^{-y^2} \\ &= \frac{\mathcal{C}}{(2\pi)^{(n-1)/2}} \left( \frac{n\Delta^2}{2} \right)^{\frac{3-n}{2}} \Gamma\left(\frac{n-3}{2}\right) \end{aligned}$$

Here we assumed that  $\sigma_{\max}^2 \gg n\Delta^2$  in the integration limits.

└ example: estimating the mean and variance

Final posterior.

$$P(\theta, \sigma^2 | \mathbf{d}) = \frac{1}{\sqrt{2\pi}\Gamma(\frac{n-3}{2})} \left(\frac{\Delta^2}{2}\right)^{\frac{n-3}{2}} \left(\frac{n}{\sigma^2}\right)^{\frac{n}{2}} \frac{1}{n} \exp\left(-\frac{n\Delta^2}{2\sigma^2}\right) \exp\left(-\frac{n(\theta - \bar{d})^2}{2\sigma^2}\right)$$

mode:

$$\ln P(\theta, \sigma^2 | \mathbf{d}) = -\frac{n}{2} \ln(\sigma^2) - \frac{n}{2\sigma^2} [\Delta^2 + (\theta - \bar{d})^2] + \text{constant terms}$$

$$\begin{aligned}\frac{\partial}{\partial \theta} \ln P(\theta, \sigma^2 | \mathbf{d}) &= -\frac{n}{\sigma^2}(\theta - \bar{d}) \\ \frac{\partial}{\partial \sigma^2} \ln P(\theta, \sigma^2 | \mathbf{d}) &= \frac{n}{2\sigma^2} \left( -1 + \frac{\Delta^2}{\sigma^2} + \frac{(\theta - \bar{d})^2}{\sigma^2} \right)\end{aligned}$$

$$\hat{\theta} = \bar{d} \quad \sigma^2 = \Delta^2 = \bar{d^2} - \bar{d}^2$$

No  $(N - 1)^{-1}$ ! It is biased.

$$\bar{d} \equiv \frac{1}{n} \sum_i d_i \quad \bar{d^2} \equiv \frac{1}{n} \sum_i d_i^2.$$

└ example: estimating the mean and variance

Final posterior.

$$P(\theta, \sigma^2 | \mathbf{d}) = \frac{1}{\sqrt{2\pi}\Gamma(\frac{n-3}{2})} \left(\frac{\Delta^2}{2}\right)^{\frac{n-3}{2}} \left(\frac{n}{\sigma^2}\right)^{\frac{n}{2}} \frac{1}{n} \exp\left(-\frac{n\Delta^2}{2\sigma^2}\right) \exp\left(-\frac{n(\theta - \bar{d})^2}{2\sigma^2}\right)$$

**mode:**

$$\ln P(\theta, \sigma^2 | \mathbf{d}) = -\frac{n}{2} \ln(\sigma^2) - \frac{n}{2\sigma^2} [\Delta^2 + (\theta - \bar{d})^2] + \text{constant terms}$$

$$\frac{\partial}{\partial \theta} \ln P(\theta, \sigma^2 | \mathbf{d}) = -\frac{n}{\sigma^2} (\theta - \bar{d})$$

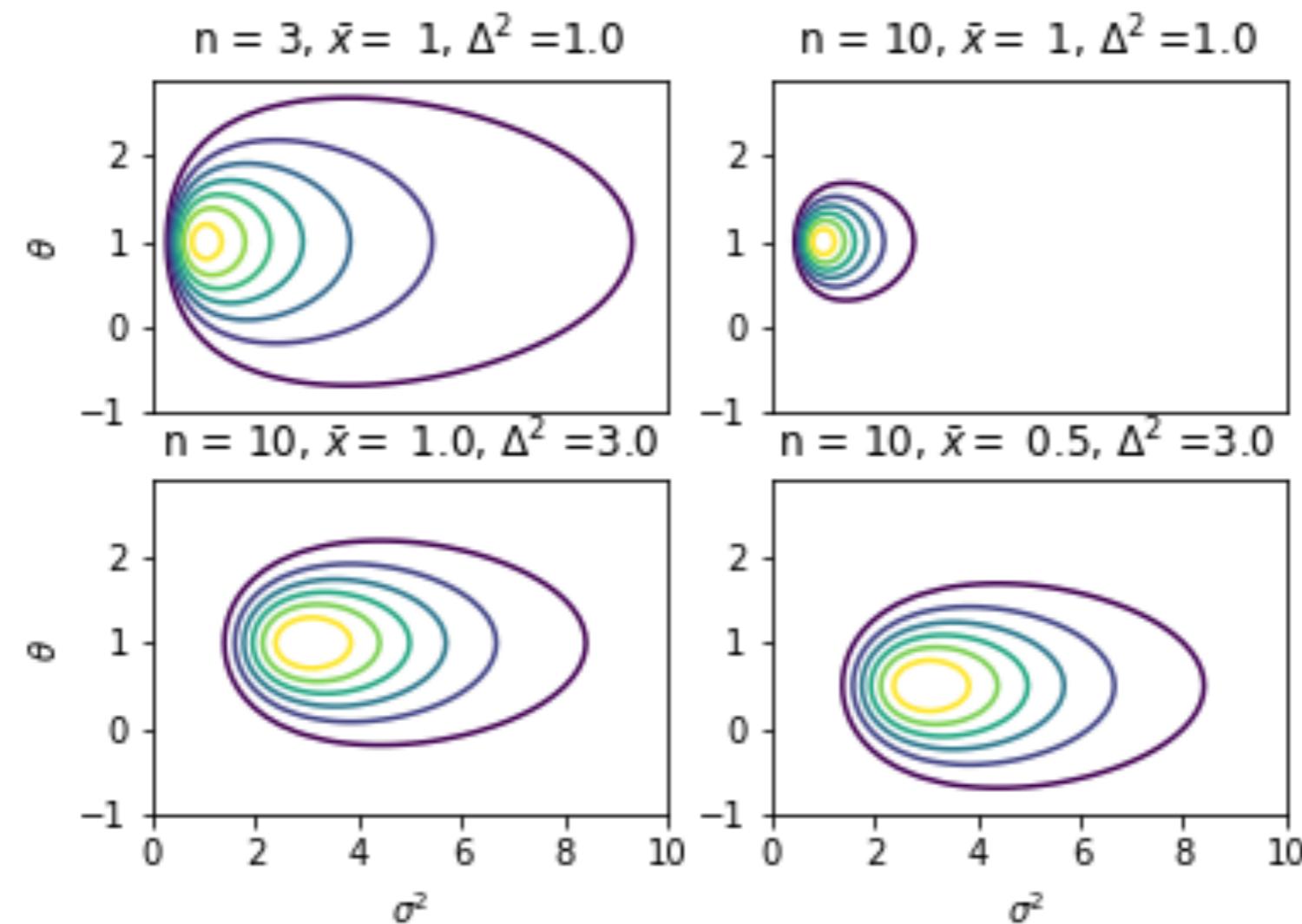
$$\frac{\partial}{\partial \sigma^2} \ln P(\theta, \sigma^2 | \mathbf{d}) = \frac{n}{2\sigma^2} \left( -1 + \frac{\Delta^2}{\sigma^2} + \frac{(\theta - \bar{d})^2}{\sigma^2} \right)$$

$$\hat{\theta} = \bar{d} \quad \hat{\sigma}^2 = \Delta^2 = \bar{d^2} - \bar{d}^2$$

No  $(N - 1)^{-1}$ ! It is biased.

$$\bar{d} \equiv \frac{1}{n} \sum_i d_i \quad \bar{d^2} \equiv \frac{1}{n} \sum_i d_i^2.$$

$$P(\theta, \sigma^2 | \mathbf{d}) = \frac{1}{\sqrt{2\pi}\Gamma\left(\frac{n-3}{2}\right)} \left(\frac{\Delta^2}{2}\right)^{\frac{n-3}{2}} \left(\frac{n}{\sigma^2}\right)^{\frac{n}{2}} \frac{1}{n} \exp\left(-\frac{n\Delta^2}{2\sigma^2}\right) \exp\left(-\frac{n(\theta - \bar{d})^2}{2\sigma^2}\right)$$



**Figure:** The posterior distributions for the mean and variance based on a sample of Gaussian distributed measurements with the number, sample mean and sample variance given above each one.

I chose to use  $\sigma^2$  as a parameter, but I could just as well have chosen  $\sigma$  or  $\sqrt{\sigma}$  as a parameter instead. The likelihoods would all be the same, but the evidence would be different since it would be an integral over a different variable. Since, by the chain rule,

$$\frac{\partial}{\partial \sigma^2} \ln P(\theta, \sigma^2 | \mathbf{d}) = \frac{1}{2\sigma} \frac{\partial}{\partial \sigma} \ln P(\theta, \sigma | \mathbf{d}) = \frac{1}{4\sigma^3} \frac{\partial}{\partial \sigma^{1/2}} \ln P(\theta, \sigma^{1/2} | \mathbf{d}) \quad (185)$$

they will all be zero at the same spot the maximum of the posterior will give the same value. However the mean parameter values will not be the same,  $\langle \sigma^2 \rangle \neq \langle \sigma \rangle^2$ .

# example: the mean without the variance

**marginalize over**  $\sigma^2$

$$\begin{aligned} P(\theta|\Delta^2, \bar{d}) &= \int_0^\infty d\sigma^2 P(\theta, \sigma^2|\Delta^2, \bar{d}) \\ &\propto \int_0^\infty d\sigma^2 \frac{e^{-\frac{A}{2\sigma^2}}}{\sigma^n} \\ &\propto -2 \int_{\infty}^0 dx x^{n-3} e^{-\frac{A}{2}x^2} \quad x = \frac{1}{\sigma} \\ &\propto 2^{\frac{n-3}{2}} A^{-\left(\frac{n-2}{2}\right)} \Gamma\left(\frac{n-2}{2}\right) \\ &\propto \left[\Delta^2 + (\theta - \bar{d})^2\right]^{\frac{2-n}{2}} \\ &\propto \left[1 + \frac{(\theta - \bar{d})^2}{\Delta^2}\right]^{\frac{2-n}{2}} \end{aligned}$$

with  $A = n\Delta^2 + n(\theta - \bar{d})^2$

$x = |\theta - \bar{d}| \sqrt{n-3} / \Delta$  is a t-distribution with  $\nu = n - 3$  degrees of freedom.

$$P(\theta|\Delta^2, \bar{d}) = \frac{\Gamma\left(\frac{n-2}{2}\right)}{\sqrt{(n-3)\pi} \Gamma\left(\frac{n-3}{2}\right)} \left[1 + \frac{(\theta - \bar{d})^2}{\Delta^2}\right]^{\frac{2-n}{2}} \quad (193)$$

Because this is symmetric about  $\bar{d}$  the mean is  $\langle \theta \rangle = \bar{d}$  and so it the mode. Since the variance of a t-distribution is  $\frac{\nu}{\nu-2}$  and  $\nu = n - 3$  in this case

$$\langle x^2 \rangle = (n-3) \frac{\langle (\theta - \bar{d})^2 \rangle}{\Delta^2} = \frac{\nu}{\nu-2} = \frac{n-3}{n-5} \quad (194)$$

so

$$\langle (\theta - \bar{d})^2 \rangle = \frac{\Delta^2}{(n-5)}. \quad (195)$$

As was discussed before, the distribution of  $t = (\bar{x} - \mu) \sqrt{n/S^2}$  is t-distributed with  $\nu = n - 1$  degrees of freedom not  $n - 3$ .

## Jeffreys prior.

$$p(\alpha) = \frac{1}{\ln(\alpha_{max}/\alpha_{min})} \begin{cases} 1/\alpha & \alpha_{min} < \alpha < \alpha_{max} \\ 0 & \text{otherwise} \end{cases} \quad (196)$$

## example: Jeffreys prior

$$P(\theta, \sigma^2 | \mathbf{d}) \propto \left( \frac{1}{\sigma^2} \right)^{\frac{1}{n}} \exp \left( -\frac{n\Delta^2}{2\sigma^2} \right) \exp \left( -\frac{n(\theta - \bar{d})^2}{2\sigma^2} \right) \quad (197)$$

By integrating this we can determine the normalization

$$P(\theta, \sigma^2 | \mathbf{d}) = \frac{n^{n/2}}{\sqrt{2^n \pi} \Gamma \left( \frac{n-1}{2} \right)} \left( \frac{\Delta^2}{2} \right)^{\frac{n-1}{2}} \frac{1}{\sigma^{n+2}} \exp \left( -\frac{n\Delta^2}{2\sigma^2} \right) \exp \left( -\frac{n(\theta - \bar{d})^2}{2\sigma^2} \right) \quad (198)$$

We can then marginalize over  $\sigma^2$  as before to get the marginalized distribution for  $\theta$

$$P(\theta | \mathbf{d}) = \frac{1}{\sqrt{\pi}} \frac{\Gamma \left( \frac{n}{2} \right)}{\Gamma \left( \frac{n-1}{2} \right)} \frac{1}{\Delta} \left[ 1 + \frac{(\theta - \bar{d})^2}{\Delta^2} \right]^{-\frac{n}{2}} \quad (199)$$

This is again a t-distribution, but now it is of  $\nu = n - 1$  degrees of freedom not  $n - 3$  degrees of freedom. Recall that  $t = (\bar{x} - \mu) \sqrt{n/s^2}$  is t-distributed with  $\nu = n - 1$  degrees of freedom.

Note also that as  $n$  gets bigger the difference between  $n - 1$  and  $n - 3$  gets less significant and the difference between the posterior distributions for uniform and Jeffreys become insignificant.

## example: Jeffreys prior radiation rate

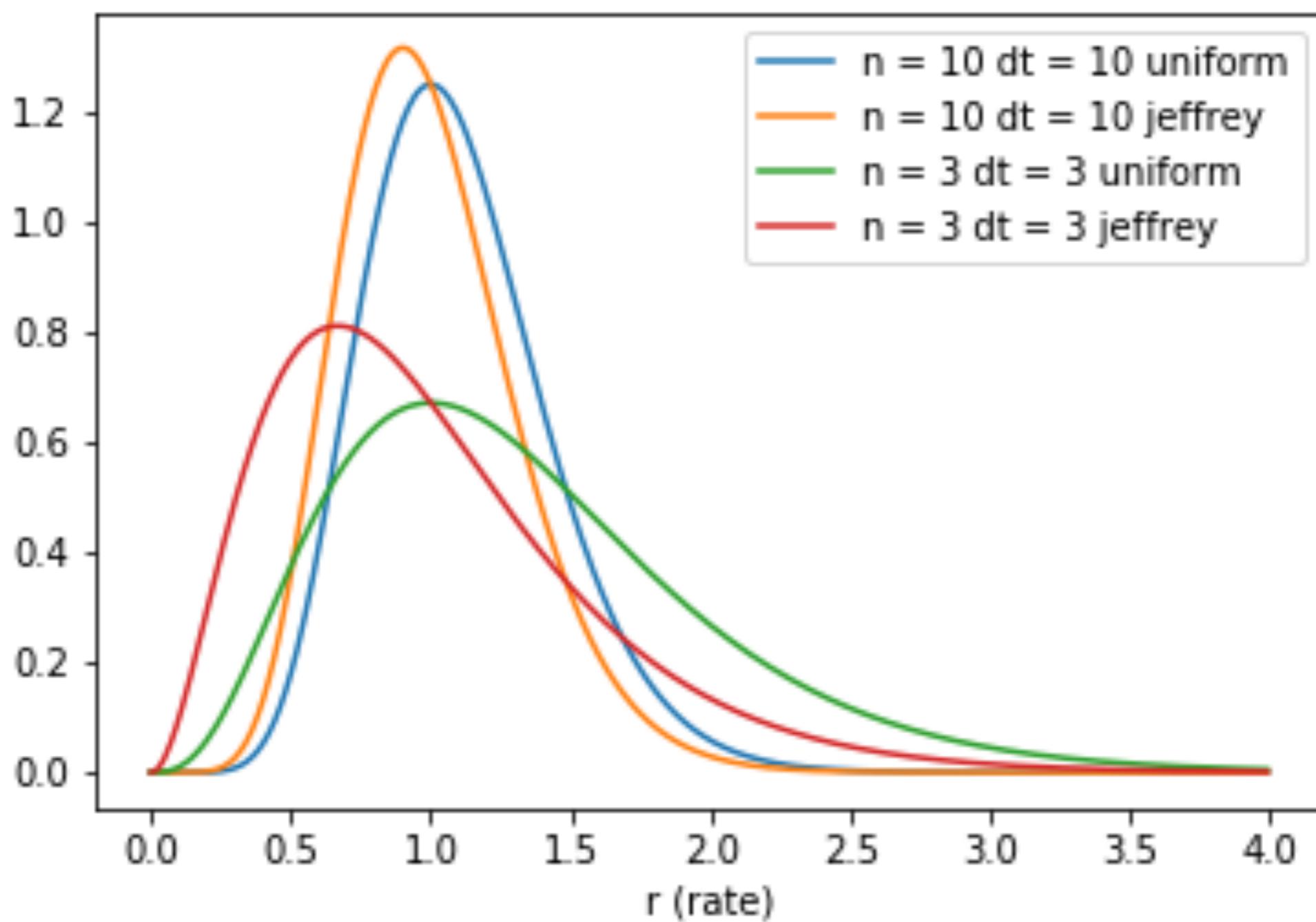
$$p(r|n) = \frac{\delta t}{(n-1)!} (\delta t r)^{n-1} e^{-r\delta t} \quad (200)$$

The mean and variance of this distribution are more in agreement with frequentist expectations

$$\langle r \rangle = \frac{n}{\delta t} \quad \text{Var}[r] = \frac{n}{\delta t^2} \quad (201)$$

Again in the limit of large  $n$  the posteriors are the same for the two choices of prior. Interestingly the maximum posterior value in this case is not  $\frac{n}{\delta t}$  but

$$r_{mode} = \frac{n-1}{\delta t} \quad (202)$$





# Example: luminosity function

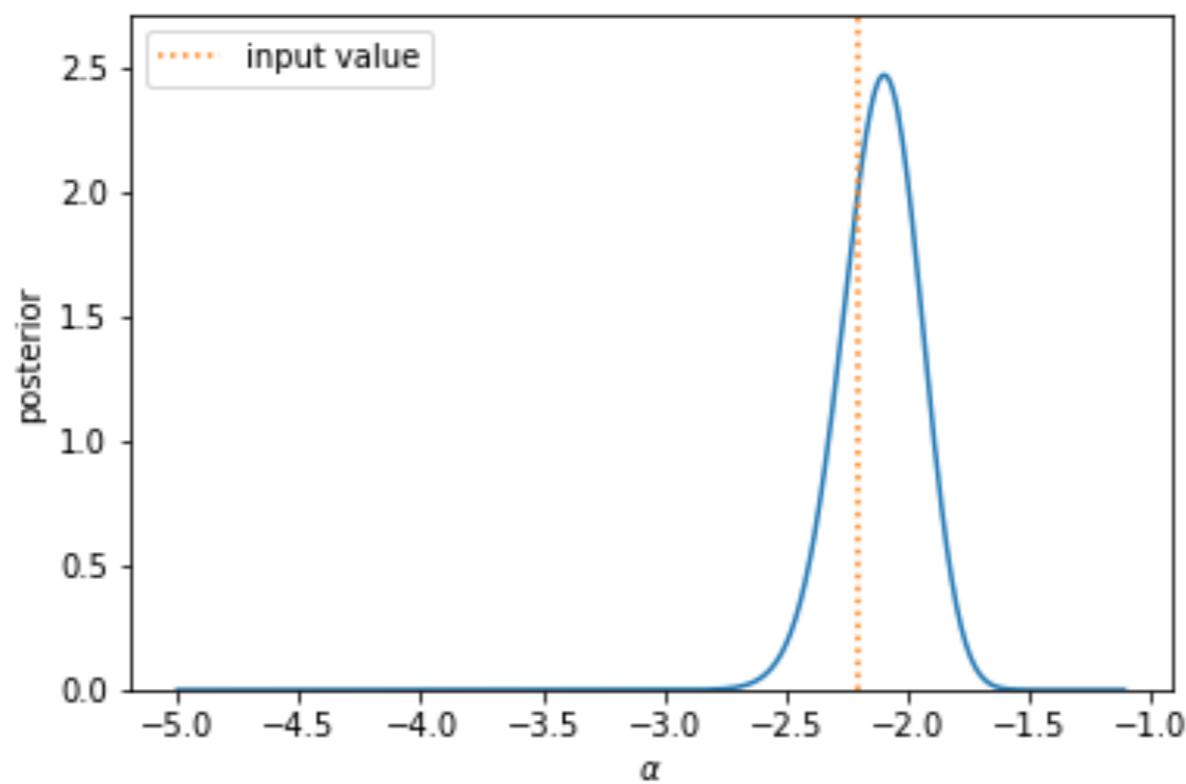
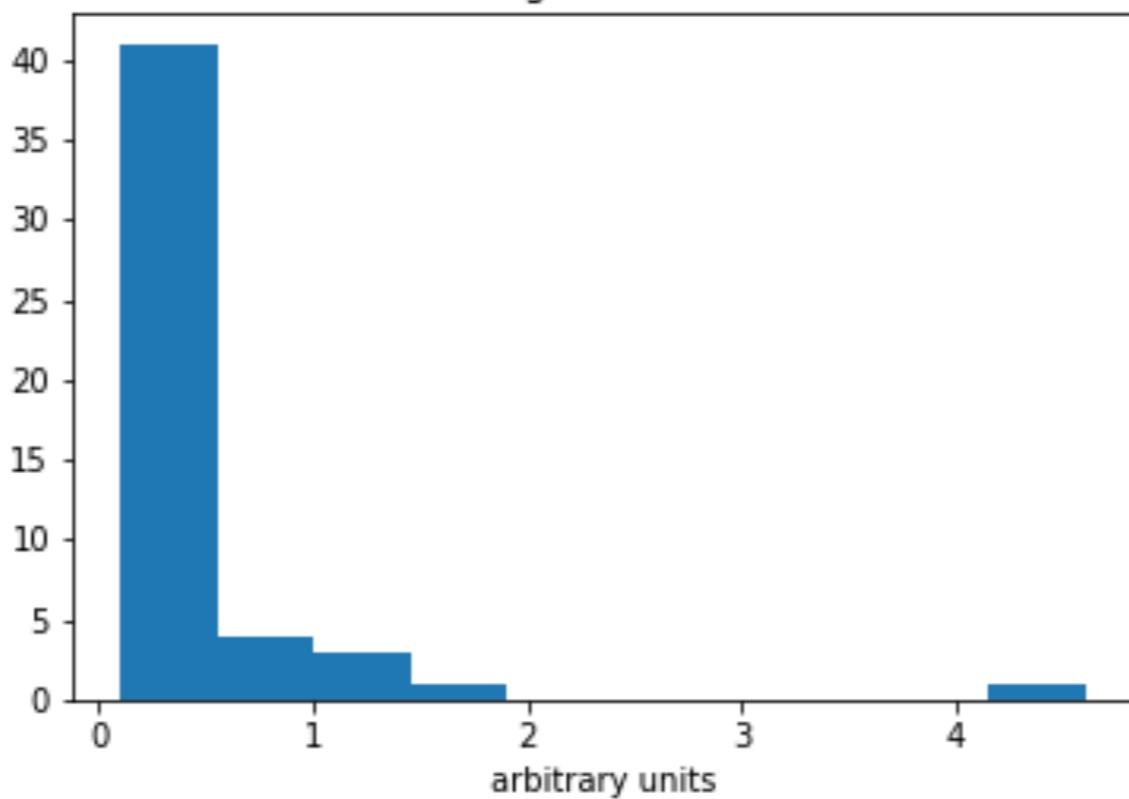
power-law luminosity function

$$f(m; \alpha) = \frac{(1 + \alpha)}{[m_{max}^{\alpha+1} - m_{min}^{\alpha+1}]} m^\alpha$$

likelihood

$$\begin{aligned}\mathcal{L}(\{m\} | \alpha) &= \prod_i f(m_i; \alpha) = \frac{(1 + \alpha)^N}{[m_{max}^{\alpha+1} - m_{min}^{\alpha+1}]^N} \prod_i m_i^\alpha \\ &= \frac{(1 + \alpha)^N}{[m_{max}^{\alpha+1} - m_{min}^{\alpha+1}]^N} \left( \prod_i m_i \right)^\alpha\end{aligned}$$

histogram of data



Maximum posterior (or likelihood) estimate of  $\alpha$

$$\ln P(\alpha | \{m_i\}) = N \ln [-(1 + \alpha)] - N \ln [m_{min}^{\alpha+1} - m_{max}^{\alpha+1}] + \alpha \ln \left( \prod_i^N m_i \right) - \ln \mathcal{E}(\{m_i\})$$

Taking  $m_{max} \rightarrow \infty$  assuming  $\alpha < -1$

$$\begin{aligned} \frac{\partial}{\partial \alpha} \ln P(\alpha | \{m_i\}) &= \frac{N}{1 + \alpha} - N \ln m_{min} + \sum_i \ln m_i \\ &= \frac{N}{1 + \alpha} + \sum_i \ln \left( \frac{m_i}{m_{min}} \right) \\ &= 0 \end{aligned}$$

$$\hat{\alpha} = - \left[ 1 + \frac{1}{\frac{1}{N} \sum_i \ln \left( \frac{m_i}{m_{min}} \right)} \right]$$

# noise in the magnitude measurement

$$p(m, m_o) = p(m)p(m_o|m)$$

where  $p(m_o|m)$  is the probability of measuring a star of magnitude  $m$  to have a magnitude  $m_o$ . Let's take this to be a Gaussian error

$$p(m_o|m) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(m_o-m)^2}{2\sigma^2}}.$$

$$\begin{aligned} p(m_o, m) &\propto f(m)e^{-\frac{(m_o-m)^2}{2\sigma^2}} \\ &= \exp \left[ \ln(f(m)) - \frac{(m_o - m)^2}{2\sigma^2} \right] \\ &= \exp \left[ \ln(f(m_o)) + \frac{\partial \ln f}{\partial m}(m - m_o) + \frac{\partial^2 \ln f}{\partial m^2}(m - m_o)^2 - \frac{(m_o - m)^2}{2\sigma^2} + \dots \right] \\ &= f(m_o) \exp \left[ \alpha \frac{(m - m_o)}{m_o} - \beta \frac{(m - m_o)^2}{m_o^2} - \frac{(m_o - m)^2}{2\sigma^2} + \dots \right] \end{aligned}$$

where

$$\alpha(m_o) \equiv \left. \frac{\partial \ln f}{\partial \ln m} \right|_{m=m_o} \quad \beta(m_o) \equiv \left( \frac{\partial \ln f}{\partial \ln m} - \frac{\partial^2 \ln f}{\partial \ln m^2} \right)_{m=\underline{m}_o}$$

# noise in the magnitude measurement

$$\begin{aligned}\frac{\partial}{\partial m} \ln p(m|m_o) &= \frac{\partial}{\partial m} \ln p(m, m_o) \\ &\simeq \frac{\alpha}{m_o} - \frac{2\beta}{m_o^2}(m - m_o) - \frac{1}{\sigma^2}(m - m_o)\end{aligned}$$

So the maximum posterior is

$$\hat{m} \simeq m_o + \frac{\alpha\sigma^2 m_o}{(m_o^2 + 2\beta\sigma^2)}$$

**Eddington bias** The observed luminosity function will be

observed luminosity function

$$f_{ob}(m_o) = p(m_o) = \int_{-\infty}^{\infty} dm \ p(m_o, m) = \int_{-\infty}^{\infty} dm \ p(m)p(m_o|m) = \int_{-\infty}^{\infty} dm \ f(m)p(m_o|m)$$

intrinsic luminosity function

$$\int_{-\infty}^{\infty} dm \ f(m)p(m_o|m)$$

↑  
errors in luminosity measurement

# intrinsic luminosity distribution

$\phi(L|\theta)$  - luminosity function

$r$  - observed distance

$R$  - true distance

$I$  - observed brightness

$L$  - true luminosity

*We want to find the luminosity function.*

*applications of the product rule*

$$\begin{aligned} p(L, I, R, r | \theta) &= p(L|\theta)p(I, R, r|L, \theta) \\ &= p(L|\theta)p(R|L, \theta)p(I, r|R, L, \theta) \\ &= p(L|\theta)p(R|L, \theta)p(I|R, L, \theta)p(r|I, R, L, \theta) \\ &= p(L|\theta)p(R|L)p(I|R, L)p(r|I, R, L) \\ &= \phi(L|\theta)p(R|L)p(I|R, L)p(r|I, R, L) \end{aligned}$$

# intrinsic luminosity distribution

$$p(R|L) \propto \frac{\partial V}{\partial R} dR \quad \text{probability of a galaxy is proportional to volume}$$

$$p(R|L)dR = 3 \left( \frac{R}{R_{\max}} \right)^2 \frac{dR}{R_{\max}} \quad \text{for Euclidian space}$$

$p(r|I, R, L)$  is the distribution of the error in the measurement of the distance  $r$ .

$$p(r|I, R, L) \simeq p((r|R)$$

$p(I|R, L)$  contains the error in the measurement of the brightness and the relationship between  $R, L$  and the brightness. If  $R$  is the luminosity distance then we might expect

$$p(I|R, L) = p \left( I - \frac{L}{4\pi R^2} \middle| \sigma_I^2 \right)$$

# intrinsic luminosity distribution

The likelihood for one object is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}|I, r) &= p(I, r|\boldsymbol{\theta}) = \int dL \int dR \, p(L, I, R, r|\boldsymbol{\theta}) \\ &= \int dL \, \phi(L|\boldsymbol{\theta}) \int dR \, p(R|L) p(I|R, L) p(r|I, R, L) \\ &\simeq \int dL \, \phi(L|\boldsymbol{\theta}) \int dR \, p(R) p\left(I - \frac{L}{4\pi R^2} \middle| \sigma_I^2\right) p(r|R) \\ &\simeq 3 \int \frac{dR}{R_{max}} \left(\frac{R}{R_{max}}\right)^2 p(r|R) \int dL \, \phi(L|\boldsymbol{\theta}) p\left(I - \frac{L}{4\pi R^2} \middle| \sigma_I^2\right)\end{aligned}$$

The posterior for the data set is then with prior  $q(\boldsymbol{\theta})$

$$p(\boldsymbol{\theta}|I, r) = \frac{\prod_i \mathcal{L}(\boldsymbol{\theta}|I_i, r_i) q(\boldsymbol{\theta})}{\int d\boldsymbol{\theta} \, [\prod_i \mathcal{L}(\boldsymbol{\theta}|I_i, r_i) q(\boldsymbol{\theta})]}$$

## selection

If  $l$  and  $r$  are observables and  $S(l, r)$  is the selection function then the likelihood for each observation is replaced with

$$\mathcal{L}_i(l_i, r_i | \theta) = p(l_i, r_i | \theta, S) = \frac{p(l_i, r_i | \theta) S(l_i, r_i)}{\int dl \int dr \ p(l, r | \theta) S(l, r)}$$

$p(l, r | \theta)$  is the probability that there is an object with properties  $l, r$  (brightness and distance in this case) and  $p(l, r | \theta, S)$  is the probability that this object will actually be observed.

$S(l, r)$  could be a magnitude limit or something more complicated that might depend on colors or surface brightness, etc.

$$S(l, r) = \begin{cases} 1 & , \quad l > l_{min} \\ 0 & , \quad l < l_{min} \end{cases}$$

Example: Simple brightness cut.

# intrinsic luminosity distribution

Idealized case where there is no noise in the distance or brightness measurements

$$\begin{aligned} p(I, r | \theta) &= \int \frac{dR}{R_{max}} 3 \left( \frac{R}{R_{max}} \right)^2 \delta^D(r - R) \int dL \phi(L | \theta) \delta^D \left( I - \frac{L}{4\pi R^2} \right) \\ &= \frac{3}{R_{max}} \left( \frac{r}{R_{max}} \right)^2 \int dL \phi(L | \theta) \delta^D \left( I - \frac{L}{4\pi r^2} \right) \\ &= \frac{3}{R_{max}} \left( \frac{r}{R_{max}} \right)^2 \int dL \phi(4\pi r^2 I | \theta) \delta^D \left( I - \frac{L}{4\pi r^2} \right) \\ &= \left( \frac{12\pi}{R_{max}^3} \right) r^4 \phi(4\pi r^2 I | \theta) \end{aligned}$$

$$\begin{aligned} \int dl \int dr p(I, r | \theta) S(I, r) &\propto \int_0^{R_{max}} dr \int_0^\infty dl r^4 \phi(4\pi r^2 I | \theta) S(I) \\ &= \int_0^{R_{max}} dr \int_{I_{min}}^\infty dl r^4 \phi(4\pi r^2 I | \theta) \\ &= \frac{1}{3(4\pi I_{max})^{3/2}} \int_0^\infty dL L^{3/2} \phi(L | \theta) \end{aligned}$$

# intrinsic luminosity distribution

The likelihood

$$\mathcal{L}(\{r_i l_i\} | \boldsymbol{\theta}) = \left(3(4\pi l_{max})^{3/2}\right)^n \frac{\prod_i r_i^4 \phi(4\pi r_i^2 l_i | \boldsymbol{\theta})}{\left[\int_0^\infty dL L^{3/2} \phi(L | \boldsymbol{\theta})\right]^n}$$

The posterior with uniform priors

$$p(\boldsymbol{\theta} | \{r_i, l_i\}) = \mathcal{C} \frac{\prod_i \phi(4\pi r_i^2 l_i | \boldsymbol{\theta})}{\left[\int_0^\infty dL L^{3/2} \phi(L | \boldsymbol{\theta})\right]^n}$$

where the normalization constant is

$$\mathcal{C}^{-1} = \int d\boldsymbol{\theta} \frac{\prod_i \phi(4\pi r_i^2 l_i | \boldsymbol{\theta})}{\left[\int_0^\infty dL L^{3/2} \phi(L | \boldsymbol{\theta})\right]^n}$$

**Malmquist bias** comes from the fact that brighter galaxies or stars sample a larger volume than dimmer ones.

# intrinsic luminosity distribution

The likelihood

$$\mathcal{L}(\{r_i l_i\} | \boldsymbol{\theta}) = \left(3(4\pi l_{max})^{3/2}\right)^n \frac{\prod_i r_i^4 \phi(4\pi r_i^2 l_i | \boldsymbol{\theta})}{\left[\int_0^\infty dL L^{3/2} \phi(L | \boldsymbol{\theta})\right]^n}$$

The posterior with uniform priors

$$p(\boldsymbol{\theta} | \{r_i, l_i\}) = \mathcal{C} \underbrace{\frac{\prod_i \phi(4\pi r_i^2 l_i | \boldsymbol{\theta})}{\left[\int_0^\infty dL L^{3/2} \phi(L | \boldsymbol{\theta})\right]^n}}_{\text{what we got before}}$$

from flux limit

where the normalization constant is

$$\mathcal{C}^{-1} = \int d\boldsymbol{\theta} \frac{\prod_i \phi(4\pi r_i^2 l_i | \boldsymbol{\theta})}{\left[\int_0^\infty dL L^{3/2} \phi(L | \boldsymbol{\theta})\right]^n}$$

**Malmquist bias** comes from the fact that brighter galaxies or stars sample a larger volume than dimmer ones.

# censoring

Censored data should not be ignored and can be incorporated into the likelihood function in a simple way. If the proposed distribution of radio fluxes is  $p(f|\theta)$  then the probability of the flux being below the threshold  $f_{max}$  is the cumulative distribution

$$F(f_{max}|\theta) = \int_{-\infty}^{f_{max}} df \ p(f|\theta)$$

so this is the factor representing an upper limit that should be used in the likelihood.



# Linear models, least-squares and regression

$$d_i = \sum_{\alpha} M_{i\alpha} \theta_{\alpha} + n_i \quad \text{or} \quad \mathbf{d} = \mathbf{M}\boldsymbol{\theta} + \mathbf{n}$$

# linear model fitting with a Gaussian likelihood

$$y = \sum_{\alpha=0}^M \theta_\alpha f_\alpha(\mathbf{x})$$

where  $y$  is the dependent variable and  $x$  is the independent variable. It is linear in the parameters  $\theta$ .

$$y_i = M_{i\alpha} \theta_\alpha \quad \text{or} \quad \mathbf{y} = \mathbf{M}\boldsymbol{\theta}$$

where the matrix  $\mathbf{M}$  contains the values of the functions  $f_\alpha(\mathbf{x})$  at each point  $\mathbf{x}_i$

$$\mathbf{M} = \begin{pmatrix} f_1(\mathbf{x}_1) & f_2(\mathbf{x}_1) & \cdots \\ f_1(\mathbf{x}_2) & f_2(\mathbf{x}_2) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

$$\ln \mathcal{L} = -\frac{1}{2} \left[ (\mathbf{y} - \mathbf{M}\boldsymbol{\theta})^T \mathbf{C}^{-1} (\mathbf{y} - \mathbf{M}\boldsymbol{\theta}) + \ln ((2\pi)^N |\mathbf{C}|) \right]$$

$$\begin{aligned}
 \frac{\partial \ln \mathcal{L}}{\partial \theta_\alpha} &= -\frac{1}{2} \frac{\partial}{\partial \theta_\alpha} \left[ (y_i - M_{i\beta} \theta_\beta) C_{ij}^{-1} (y_j - M_{j\gamma} \theta_\gamma) + \ln ((2\pi)^N |\mathbf{C}|) \right] \\
 &= \frac{1}{2} \left[ M_{i\alpha} C_{ij}^{-1} (y_j - M_{j\gamma} \theta_\gamma) + (y_i - M_{i\beta} \theta_\beta) C_{ij}^{-1} M_{j\alpha} \right] \\
 &= M_{i\alpha} C_{ij}^{-1} y_j - M_{i\alpha} C_{ij}^{-1} M_{j\gamma} \theta_\gamma \quad \mathbf{C} \text{ is symmetric} \\
 &= \mathbf{M}^T \mathbf{C}^{-1} \mathbf{y} - \mathbf{M}^T \mathbf{C}^{-1} \mathbf{M} \boldsymbol{\theta}
 \end{aligned}$$

Maximum Likelihood Estimator (MLE)

$$\hat{\boldsymbol{\theta}} = \left( \mathbf{M}^T \mathbf{C}^{-1} \mathbf{M} \right)^{-1} \mathbf{M}^T \mathbf{C}^{-1} \mathbf{y}$$

The posterior is

$$p(\theta | \mathbf{y}, \mathbf{x}) = \frac{\sqrt{|\mathbf{M}^T \mathbf{C}^{-1} \mathbf{M}|}}{(2\pi)^{N/2}} \exp \left[ -\frac{1}{2} (\theta - \hat{\theta})^T \mathbf{M}^T \mathbf{C}^{-1} \mathbf{M} (\theta - \hat{\theta}) \right]$$

The covariance for the *parameters* is  $(\mathbf{M}^T \mathbf{C}^{-1} \mathbf{M})^{-1}$ .

# fitting a line

$$y = \theta_0 + \theta_1 x$$

$$\begin{aligned}\mathbf{M}^T \mathbf{C}^{-1} \mathbf{M} &= \frac{1}{\sigma^2} \begin{pmatrix} 1 & 1 & 1 & \dots \\ x_1 & x_2 & x_3 & \dots \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \end{pmatrix} \\ &= \frac{1}{\sigma^2} \begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \\ &= \frac{N}{\sigma^2} \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \bar{x}^2 \end{pmatrix}\end{aligned}$$

$$(\mathbf{M}^T \mathbf{C}^{-1} \mathbf{M})^{-1} = \frac{\sigma^2}{N(\bar{x}^2 - \bar{x}^2)} \begin{pmatrix} \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

Best fit or maximum posterior solution

$$\hat{\theta} = (\mathbf{M}^T \mathbf{C}^{-1} \mathbf{M})^{-1} \mathbf{M}^T \mathbf{C}^{-1} \mathbf{y}$$

$$= \frac{1}{N(\bar{x}^2 - \bar{x}^2)} \begin{pmatrix} \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & \dots \\ x_1 & x_2 & x_3 & \dots \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{pmatrix}$$

$$= \frac{1}{N(\bar{x}^2 - \bar{x}^2)} \begin{pmatrix} \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix}$$

$$= \frac{1}{(\bar{x}^2 - \bar{x}^2)} \begin{pmatrix} \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} \bar{y} \\ \bar{xy} \end{pmatrix}$$

$$= \frac{1}{(\bar{x}^2 - \bar{x}^2)} \begin{pmatrix} \bar{x}^2 \bar{y} - \bar{x} \bar{xy} \\ \bar{xy} - \bar{x} \bar{y} \end{pmatrix}$$

# when both variables are uncertain

$$\begin{aligned}\mathcal{L}(\mathbf{x}^o, \mathbf{y}^o | \boldsymbol{\theta}, \mathbf{x}) &= \frac{1}{(2\pi\sigma_x^2\sigma_y^2)^{N/2}} \exp\left[-\frac{1}{2} \sum_i \frac{(y_i^o - \theta_0 - \theta_1 x_i)^2}{\sigma_y^2}\right] \exp\left[-\frac{1}{2} \sum_i \frac{(x_i^o - x_i)^2}{\sigma_x^2}\right] \\ &= \prod_i \mathcal{G}(y_i^o | \theta_0 + \theta_1 x_i, \sigma_y^2) \mathcal{G}(x_i^o | x_i, \sigma_x^2) \\ &= \prod_i \mathcal{G}(\theta_1 x_i | y_i^o - \theta_0, \sigma_y^2) \mathcal{G}(x_i | x_i^o, \sigma_x^2) \\ &= \prod_i \frac{1}{\theta_1} \mathcal{G}\left(x_i \left| \frac{y_i^o - \theta_0}{\theta_1}, \frac{\sigma_y^2}{\theta_1^2}\right.\right) \mathcal{G}(x_i | x_i^o, \sigma_x^2) \\ &= \frac{1}{\theta_1^N} \prod_i \mathcal{G}\left(\frac{y_i^o - \theta_0}{\theta_1} \left| x_i^o, \sigma_x^2 + \frac{\sigma_y^2}{\theta_1^2}\right.\right) \mathcal{G}\left(x_i \left| \mu_c, \sigma_x^2 + \frac{\sigma_y^2}{\theta_1^2}\right.\right)\end{aligned}$$

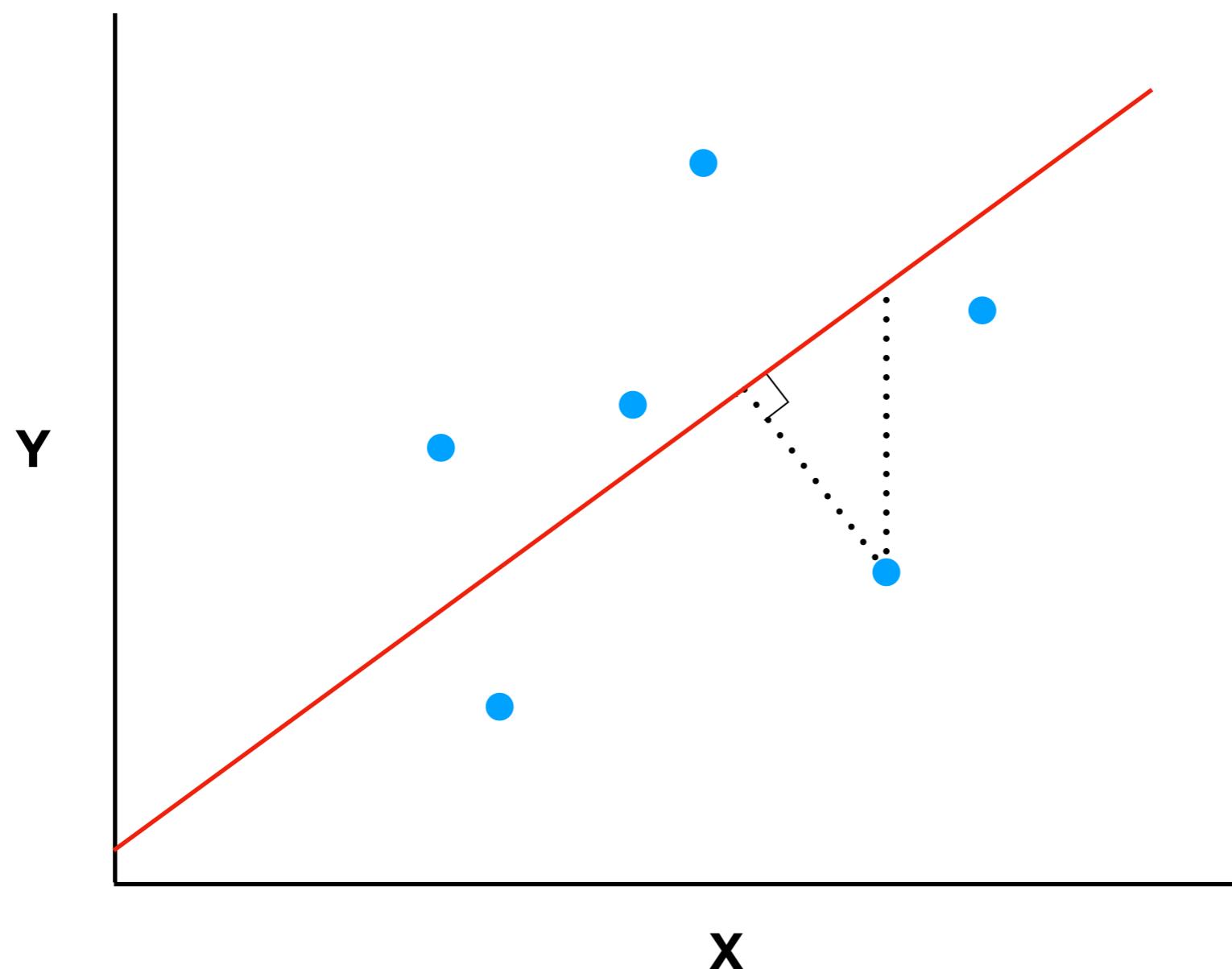
## Marginalized posterior

$$\begin{aligned}
 P(\boldsymbol{\theta} | \mathbf{x}^o, \mathbf{y}^o) &= \int d^n x \, P(\boldsymbol{\theta}, \mathbf{x} | \mathbf{x}^o, \mathbf{y}^o) \\
 &= \frac{C}{(2\pi(\sigma_y^2 + \theta_1^2 \sigma_x^2))^{N/2}} \exp \left[ -\frac{1}{2} \sum_i \frac{(y_i^o - \theta_0 - \theta_1 x_i^o)^2}{(\sigma_y^2 + \theta_1^2 \sigma_x^2)} \right]
 \end{aligned}$$

## Maximum posterior (likelihood) estimators

$$\hat{\theta}_0 = \frac{\bar{xy} - \bar{x} \bar{y}}{(\sigma_x^2 + \bar{x^2} - \bar{x}^2)}$$

$$\hat{\theta}_1 = \frac{\bar{y} \bar{x^2} + \bar{y} \sigma_x^2 - \bar{xy} \bar{x}}{(\sigma_x^2 + \bar{x^2} - \bar{x}^2)}$$



# regression with censored data

For the case of the a linear model and Gaussian errors the likelihood for each upper limit is

$$\begin{aligned} F(y_{\text{upper}}|x) &= \int_{-\infty}^{y_{\text{upper}}} dy' p\left(y' \middle| \sum_{\alpha} \theta_{\alpha} f_{\alpha}(x) \sigma\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{y_{\text{upper}}} dy' \exp\left[-\frac{1}{2\sigma^2} \left(y' - \sum_{\alpha} \theta_{\alpha} f_{\alpha}(x)\right)^2\right] \\ &= \frac{1}{2} \left[ 1 + \operatorname{erf}\left(\frac{y_{\text{upper}} - \sum_{\alpha} \theta_{\alpha} f_{\alpha}(x)}{\sigma}\right) \right] \end{aligned}$$

# Least-Squares

$$M_{SE} = \|\mathbf{y} - \mathbf{M}\boldsymbol{\theta}\|_2^2 \equiv \sum_i \left( y_i - \sum_{\alpha} M_{i\alpha} \theta_{\alpha} \right)^2$$

**mean squared error** or MSE. (It is conventional to define  
 $\|\mathbf{x}\|_p \equiv (\sum_i x_i^p)^{1/p}$ . )

$$\hat{\boldsymbol{\theta}}_{LS} = (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T \mathbf{y}$$

The matrix  $(\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T$  is sometimes called the **pseudoinverse** or **Moore-Penrose inverse** of the matrix  $\mathbf{M}$  (replacing the transposes with the Hermitian transpose for complex matrices).



# Matrix decompositions

## Eigenvalue decomposition

If  $A$  is a  $N \times N$  matrix with linear independent columns the it can be decomposed as

$$A = M\Lambda M^{-1}$$

where  $\Lambda$  is diagonal and  $\Lambda_{ii}$  is the  $i$ th eigenvalue and the  $i$ th column of  $M$  is the corresponding eigenvalue.

## Single-value decomposition

If  $A$  is a  $M \times N$  matrix it can be factorized as

$$A = SVD^\dagger$$

where

- $S$  is a unitary (orthogonal if real)  $M \times M$  matrix, i.e.  $SS^\dagger = S^\dagger S = I$
- $V$  is a diagonal matrix  $M \times N$  with non-negative real entries
- $D$  is a unitary(orthogonal if real)  $N \times N$  matrix

# calculating the pseudoinverse

The SVD decomposition of  $\mathbf{M}$  is  $\mathbf{M} = \mathbf{S}\mathbf{V}\mathbf{D}^T$ , where  $\mathbf{V}$  is a diagonal matrix, but it is not square. The number of columns will be the number of parameters and the number of rows will be the number of data points. The pseudoinverse of  $\mathbf{M}$  is then

$$\mathbf{M}^+ = \mathbf{D}\mathbf{V}^+\mathbf{S}^T$$

where  $\mathbf{V}^+$  is found by taking the reciprocal of the nonzero entries, i.e.  $V_{ii}^+ = 1/V_{ii}$  for  $V_{ii} \neq 0$ .

Computational time  $\mathcal{O} [\min(m^2n, mn^2)]$ .

The minimum  $\chi^2$  problem can be converted into a least-squares problem by **pre-whitening** the data.

# Bayesian Prediction

Predict the value  $y$  at a new point  $x$ .  $\mathbf{Y}, \mathbf{X}$  is the "training set".

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}, \mathbf{Y}, \mathbf{X}) &= \int_{-\infty}^{\infty} d\theta \ p(\mathbf{y}\theta|\mathbf{x}, \mathbf{Y}, \mathbf{X}) \\ &= \int_{-\infty}^{\infty} d\theta \ p(\mathbf{y}|\theta, \mathbf{x}, \mathbf{Y}, \mathbf{X}) p(\theta|\mathbf{x}, \mathbf{Y}, \mathbf{X}) \\ &= \int_{-\infty}^{\infty} d\theta \ p(\mathbf{y}|\theta, \mathbf{x}) p(\theta|\mathbf{Y}, \mathbf{X}). \end{aligned}$$

$$p(\mathbf{y}|\mathbf{x}, \mathbf{Y}, \mathbf{X}) = \int_{-\infty}^{\infty} d\theta \ \delta^D (\mathbf{y} - \mathbf{f}(\mathbf{x}|\theta)) p(\theta|\mathbf{Y}, \mathbf{X})$$

where  $\mathbf{f}(\mathbf{x}|\theta)$  is the regression function.

# Bayesian Prediction

Predict the value  $y$  at a new point  $x$ .  $\mathbf{Y}, \mathbf{X}$  is the "training set".

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}, \mathbf{Y}, \mathbf{X}) &= \int_{-\infty}^{\infty} d\theta \ p(\mathbf{y}|\theta, \mathbf{x}, \mathbf{Y}, \mathbf{X}) \\ &= \int_{-\infty}^{\infty} d\theta \ p(\mathbf{y}|\theta, \mathbf{x}, \mathbf{Y}, \mathbf{X}) p(\theta|\mathbf{x}, \mathbf{Y}, \mathbf{X}) \\ &= \int_{-\infty}^{\infty} d\theta \ p(\mathbf{y}|\theta, \mathbf{x}) p(\theta|\mathbf{Y}, \mathbf{X}). \end{aligned}$$

*prediction given parameters*

*posterior given training set*

$$p(\mathbf{y}|\mathbf{x}, \mathbf{Y}, \mathbf{X}) = \int_{-\infty}^{\infty} d\theta \ \delta^D (\mathbf{y} - \mathbf{f}(\mathbf{x}|\theta)) p(\theta|\mathbf{Y}, \mathbf{X})$$

where  $\mathbf{f}(\mathbf{x}|\theta)$  is the regression function.

# nonparametric regression

kernel function  $K(x)$ . It might be a top-hat function, a Gaussian, B-spline or something else. It is symmetric about  $x = 0$  and drops off rapidly for  $|x| > 1$ . The estimated function  $\hat{f}(x)$  is then

$$\hat{f}(x) = \frac{\sum_{i=1}^n y_i K\left(\frac{x-x_i}{h_x}\right)}{\sum_{i=1}^n K\left(\frac{x-x_i}{h_x}\right)}$$

where  $h_x$  is a scale factor that needs to be chosen.

The average of this is an unbiased estimator of the kernel smoothed function

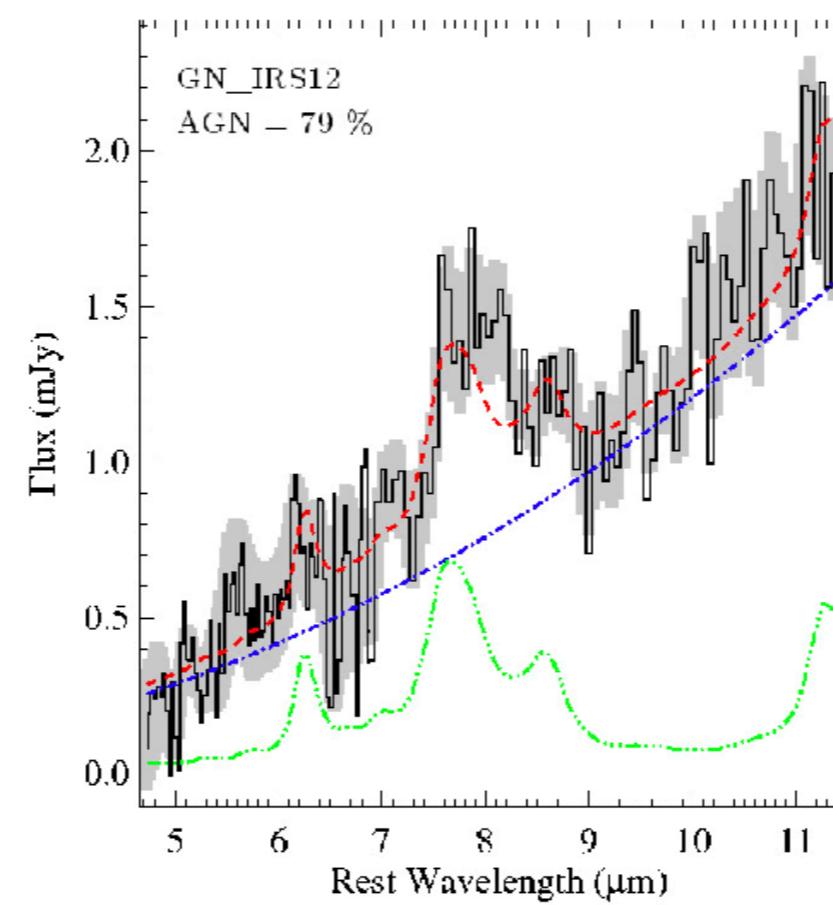
$$\langle \hat{f}(x) \rangle = \frac{\sum_{i=1}^n f(x_i) K\left(\frac{x-x_i}{h_x}\right)}{\sum_{i=1}^n K\left(\frac{x-x_i}{h_x}\right)}.$$

and the variance if the noise is uncorrelated between data points is

$$\begin{aligned} \sigma_{\hat{f}(x)}^2 &= \langle \hat{f}(x)^2 \rangle - \langle \hat{f}(x) \rangle^2 \\ &= \frac{\sum_{i=1}^n \sigma_i^2 K\left(\frac{x-x_i}{h_x}\right)^2}{\left[\sum_{i=1}^n K\left(\frac{x-x_i}{h_x}\right)\right]^2}. \end{aligned}$$

20/06/2019

AGN\_example\_fit1.png (567x567)



# bootstrap (nonparametric bootstrap)

Try to recover the variance of a statistics from a limited sample.

**bootstrap PDF**

$$f^{bs}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^N \delta(\mathbf{x} - \mathbf{x}_i)$$

This can be considered the maximum likelihood estimate for the pdf of the data itself.  
It is an estimate for the real pdf of the data.

Let's consider any statistic that is a function of these data point  $t(x_1, x_2, \dots, x_n)$ .

Assuming that each data point is statistically independent, the expectation value of this statistic will be

$$E[t(x_1, x_2, \dots, x_n)] = \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_n p(x_1) \dots p(x_n) t(x_1, x_2, \dots, x_n).$$

Using the bootstrap estimation of the pdf (334) gives

$$\begin{aligned} E^{bs}[t(x_1, x_2, \dots, x_n)] &= \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_n f^{bs}(x_1) \dots f^{bs}(x_n) t(x_1, x_2, \dots, x_n) \\ &= \frac{1}{n^n} \sum_{i_1=1}^n \dots \sum_{i_n=1}^n t(x_{i_1}, x_{i_2}, \dots, x_{i_n}) \end{aligned}$$

These sums contain all possible combinations of the data in the "slots" of  $t(x_{i_1}, x_{i_2}, \dots, x_{i_n})$ . All of these combinations except one, the original data, has repeated values in them.

There are

$$\binom{2n-1}{n}$$

distinct bootstrap samples which is 92378 for  $n = 10$  and  $\simeq 4.53 \times 10^{58}$  for  $n = 100$

Approximation:

- Resample from the data **with replacement** a new data set of the same size.
- Calculate the statistic,  $t$ .
- repeat for  $N_{boot} \gtrsim 1000$
- find the variance of the  $N_{boot}$  using the usual estimator:

$$Var^{bs}[t] = \frac{1}{N_{boot} - 1} \sum_{i=1}^{N_{boot}} (t_i - \bar{t}_{bs})^2$$

$$\bar{t}_{bs} = \frac{1}{N_{boot}} \sum_{i=1}^{N_{boot}} t_i$$

**DEMO**

**bootstrap\_regression.py**

# jackknife sampling

$t_n$  is a statistic based on  $n$  independent data points

$$t_n = t_\infty + \frac{t_b}{n} + \mathcal{O}(n^{-2})$$

Applying this to the  $n - 1$  case gives

$$t_{n-1} = t_\infty + \frac{t_b}{n-1} + \mathcal{O}(n^{-2})$$

Combining these we can eliminate the lowest order bias and solve for the asymptotic limit

$$\begin{aligned} t_\infty &= nt_n + (1-n)t_{n-1} + \mathcal{O}(n^{-2}) \\ &= t_n + (n-1)(t_n - t_{n-1}) + \mathcal{O}(n^{-2}) \end{aligned}$$

$$\bar{t}_{n-1}^J \equiv \frac{1}{n} \sum_{i=1}^n t_{n-1}^{(i)}$$

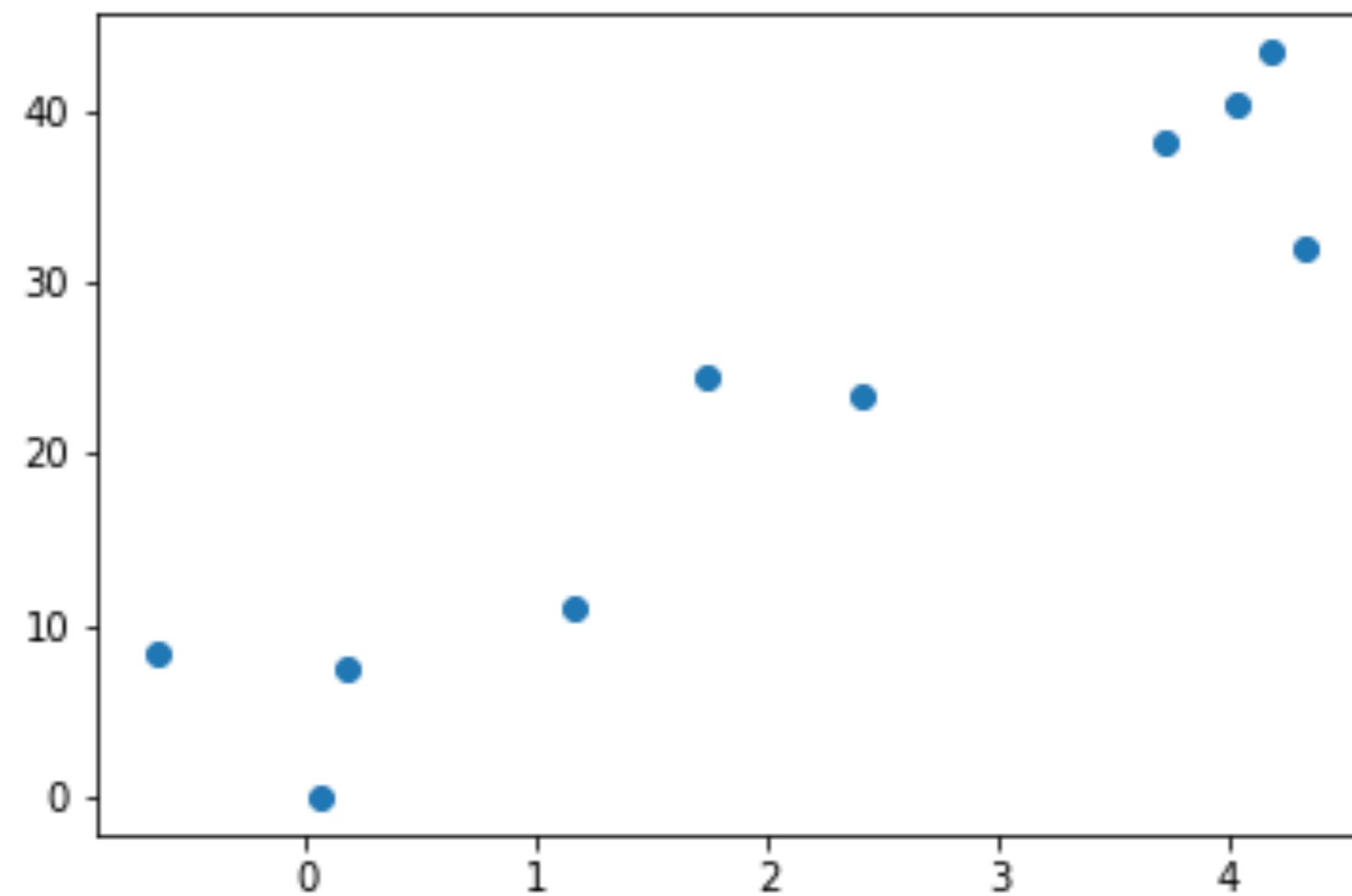
The jackknife estimate of the statistic  $t$

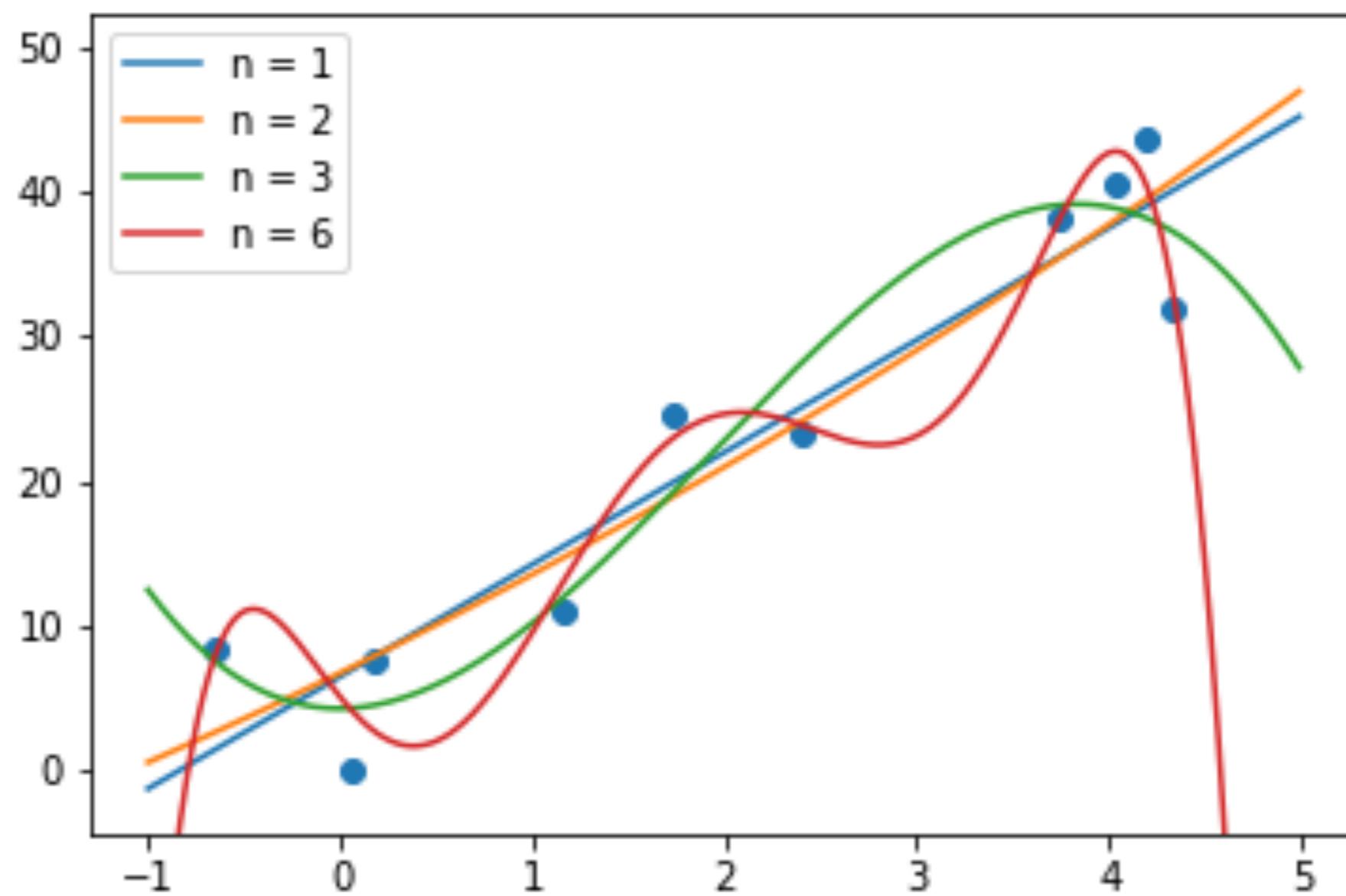
$$t_n^J = nt_n + (1 - n)\bar{t}_{n-1}^J$$

which includes a correction for bias. We can also calculate the jackknife estimate of

the variance variance for our statistic

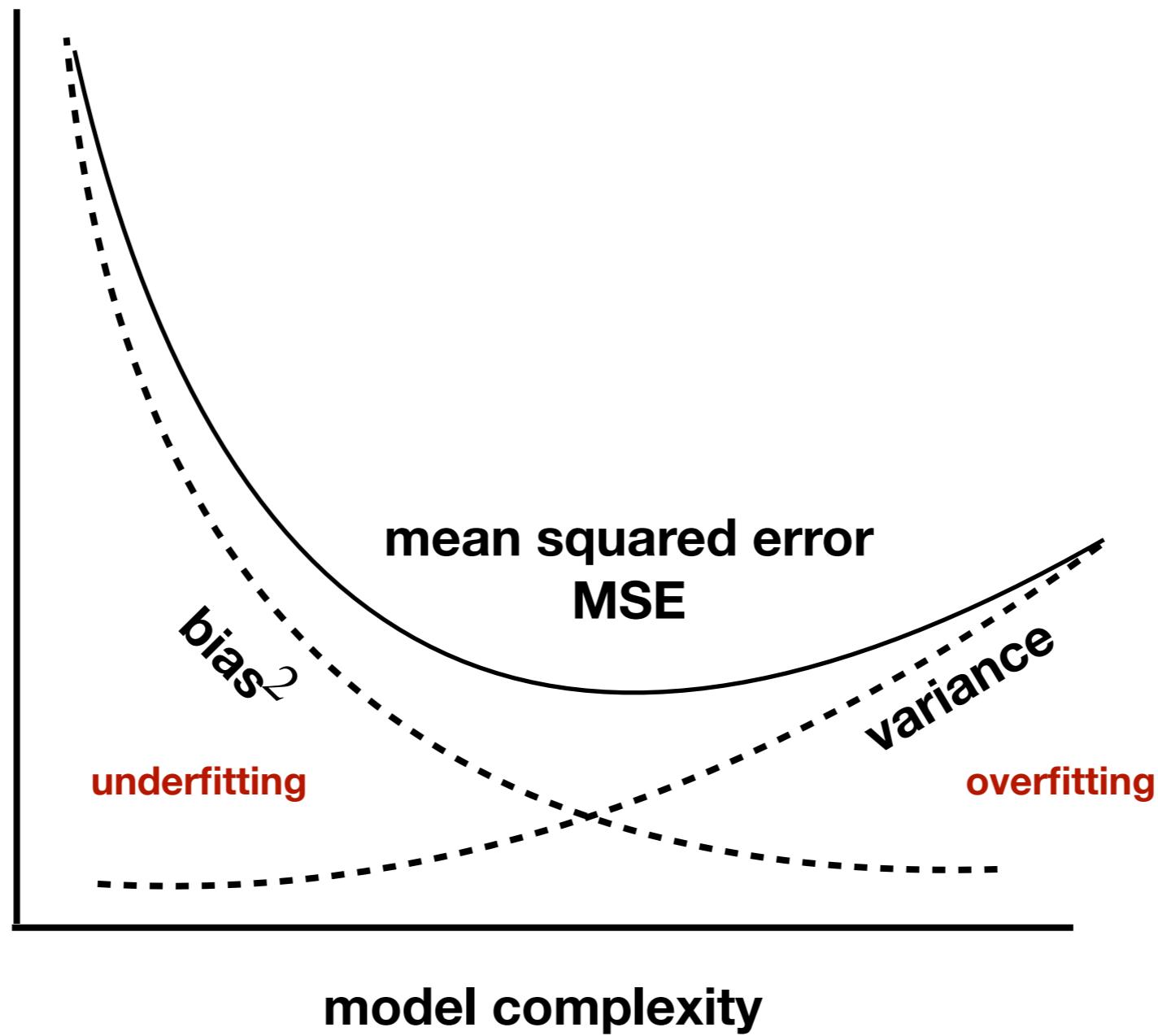
$$Var^J[t_n] = \frac{n-1}{n} \sum_{i=1}^n \left( t_{n-1}^{(i)} - \bar{t}_{n-1}^J \right)^2$$





# bias-variance trade-off

$$\begin{aligned}\langle (y - \hat{f}_{\{x_i\}}(x))^2 \rangle &= \left\langle (f(x) + n_y - \hat{f}_{\{x_i\}}(x))^2 \right\rangle \\&= \left\langle f(x)^2 + n_y^2 + \hat{f}_{\{x_i\}}(x)^2 - 2f(x)\hat{f}_{\{x_i\}}(x) \right\rangle \quad \langle n_y \rangle = 0 \\&= f(x)^2 + \sigma_y^2 + \left\langle \hat{f}_{\{x_i\}}(x)^2 \right\rangle - 2f(x)\left\langle \hat{f}_{\{x_i\}}(x) \right\rangle \\&= f(x)^2 + \sigma_y^2 + \text{Var}[\hat{f}_{\{x_i\}}(x)] + \left\langle \hat{f}_{\{x_i\}}(x) \right\rangle^2 - 2f(x)\left\langle \hat{f}_{\{x_i\}}(x) \right\rangle \\&= \sigma_y^2 + \text{Var}[\hat{f}_{\{x_i\}}(x)] + \left( f(x) - \left\langle \hat{f}_{\{x_i\}}(x) \right\rangle \right)^2 \\&= \sigma_y^2 + \text{Var}[\hat{f}_{\{x_i\}}(x)] + \text{Bias} [\hat{f}_{\{x_i\}}(x)]^2\end{aligned}$$



# cross-validation

## *k*-fold cross-validation.

- The data is split into  $k$  subsets.
- $k - 1$  of the subsets are the *training* set.
- The remaining subset that was not used in the fit is called the *validation* set.
- If we call the model fit to all but the  $j$ th set  $\hat{Y}_{-j}(x)$  then the predicted error is

$$PE = \frac{1}{k} \sum_{j=1}^k \frac{1}{n_j} \sum_{i \in \{j\}} [y_i - \hat{Y}_{-j}(x_i)]^2 \quad \text{predicted error}$$

where the inner sum is over the subset left out.

- The number of parameters can be increased until the PE reaches a minimum and starts to increasing due to over fitting.
- Finally the model parameters for each training models are averaged.
- A measure of the bias can be found by subtracting the MSE from the whole set from the PE.

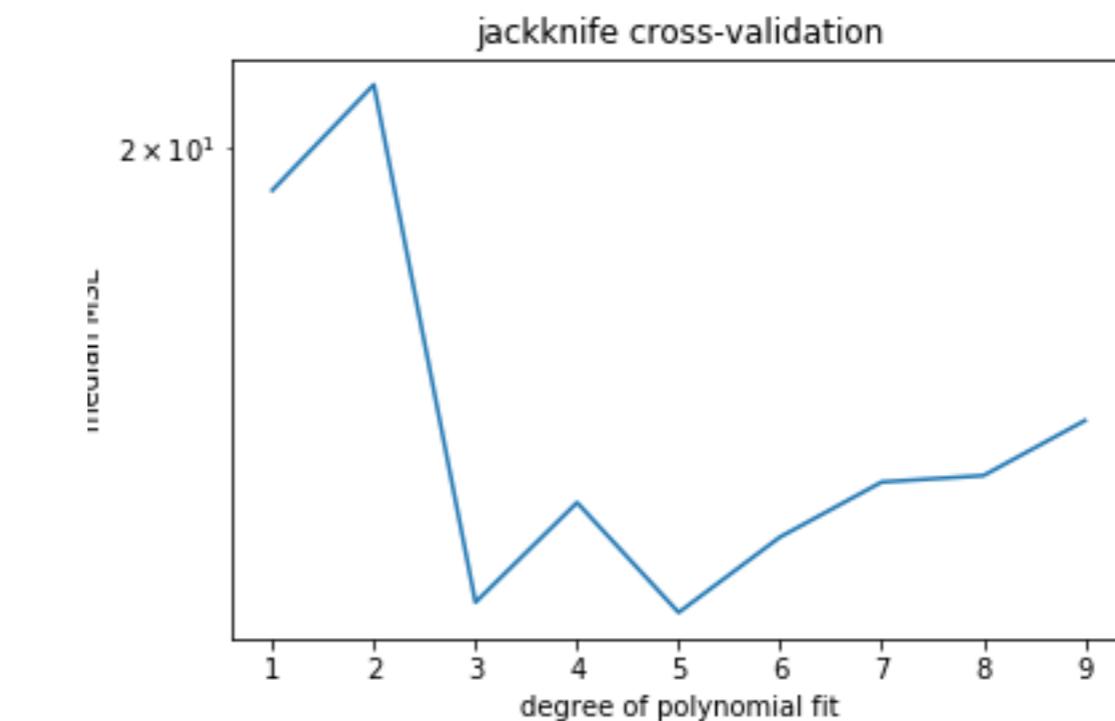
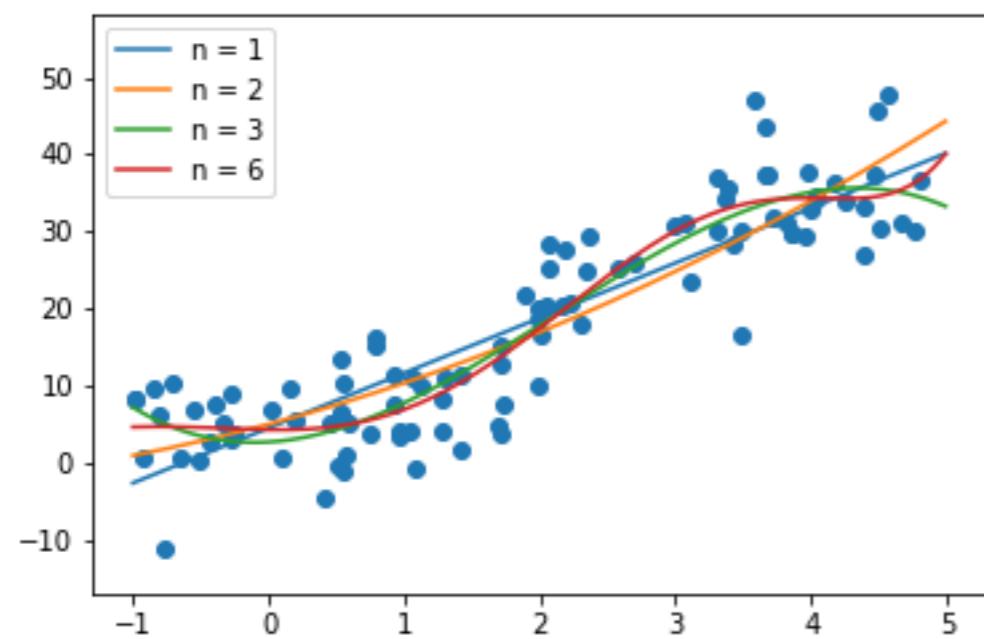
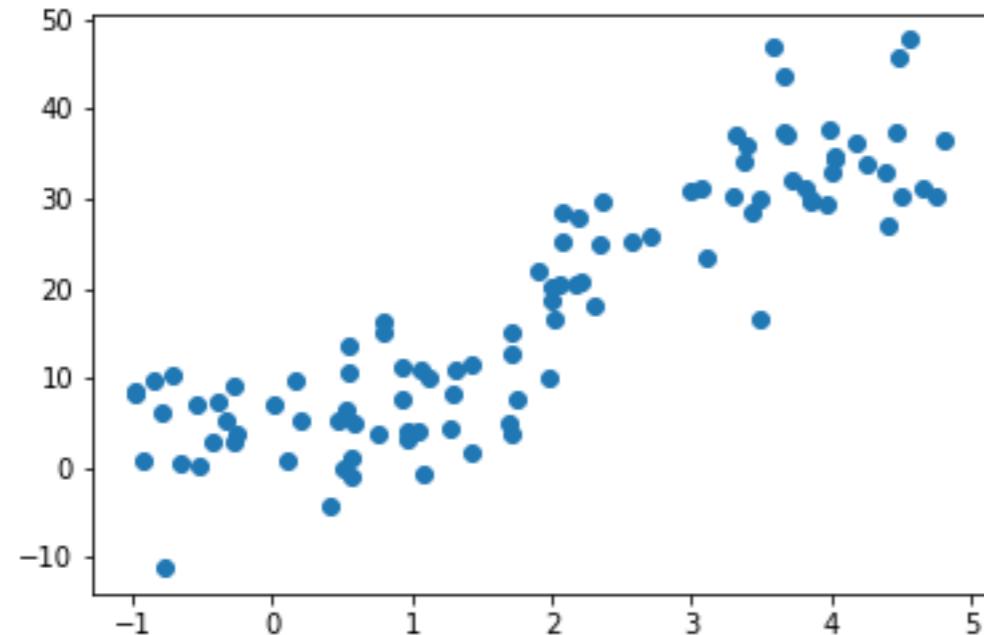
**DEMO**

**k-fold-demo.py**

**DEMO**

**poly\_fits.py**

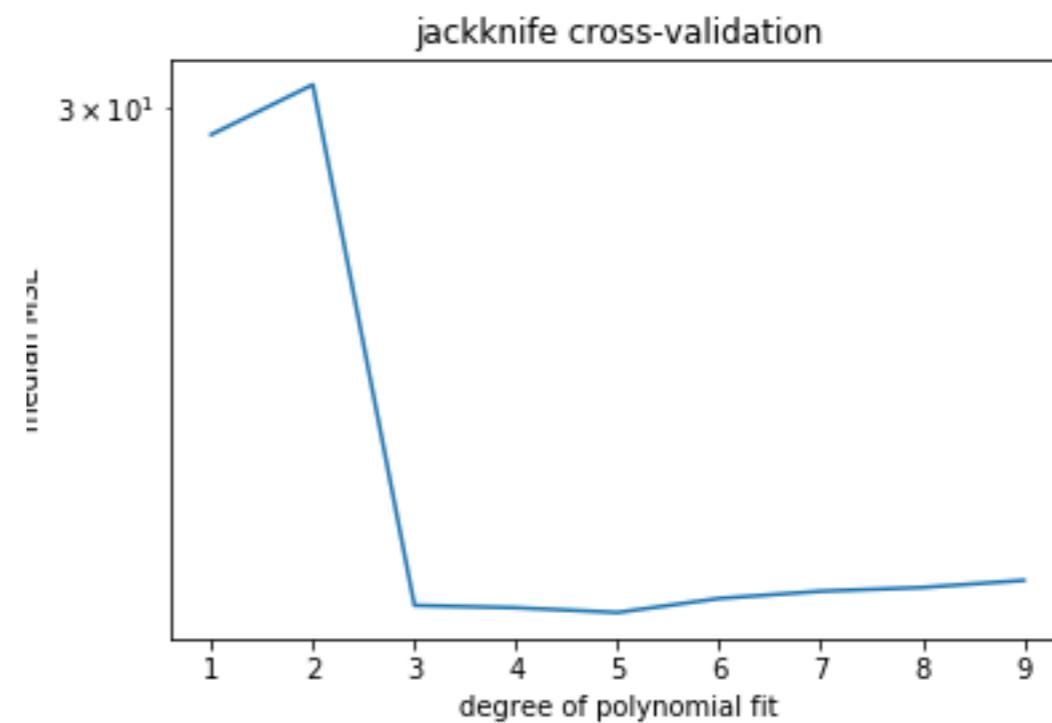
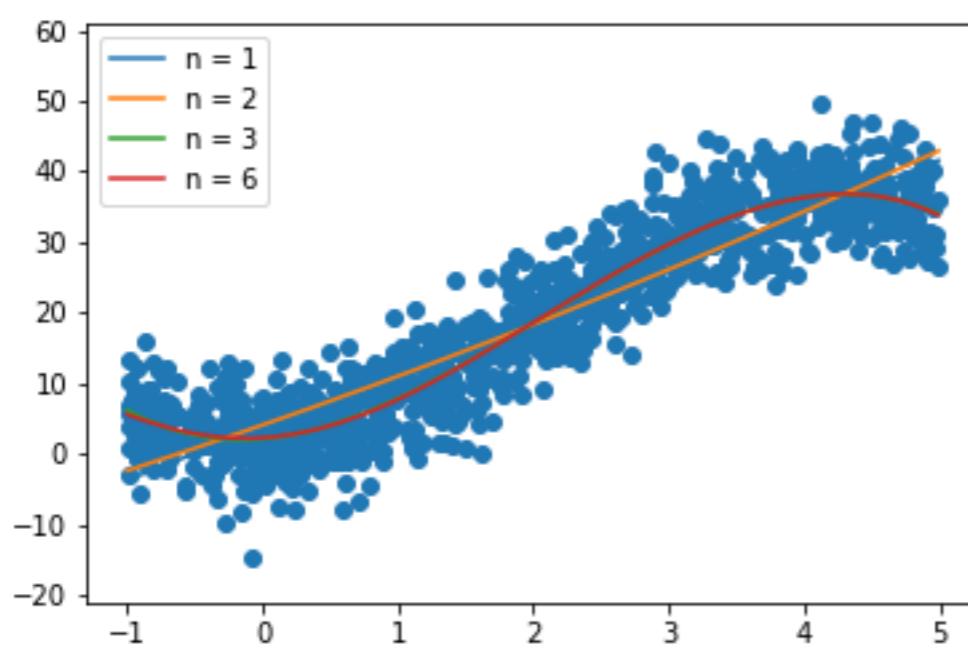
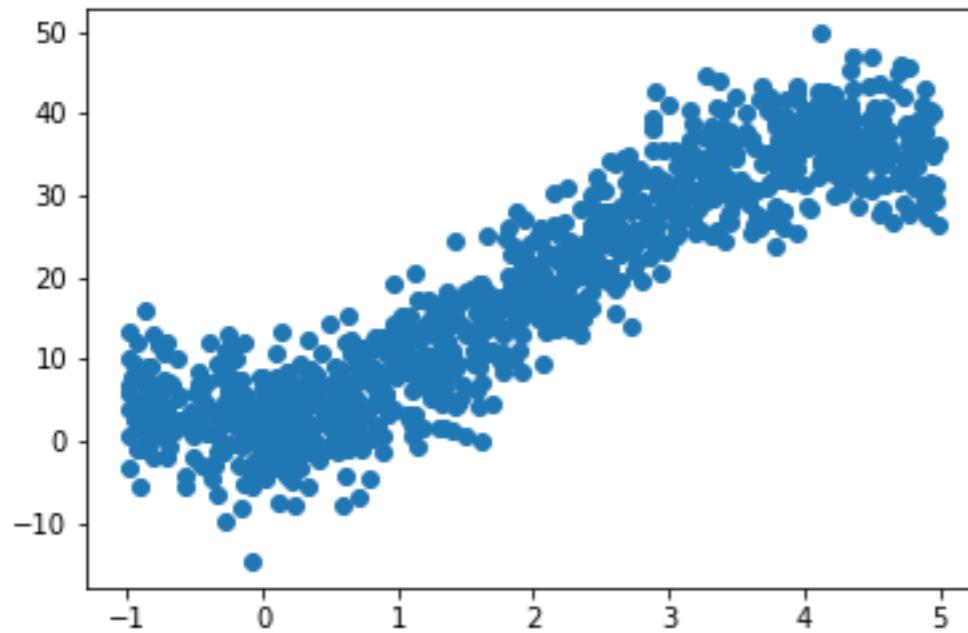
# Model Selection with Cross-Validation



Bootstrap parameter estimation :

p[3] -0.7462223047294956 +/- 0.17348500621532936  
p[2] 4.716878889007931 +/- 0.9759334581030513  
p[1] 1.143767821401198 +/- 1.354971733358651  
y intercept 2.7391260701456708 +/- 0.9711388573262643

# Model Selection with Cross-Validation



Bootstrap parameter estimation :

p[3] -0.7872804116968187 +/- 0.04201237033437874  
p[2] 4.885536563461591 +/- 0.25914690571999144  
p[1] 1.6136524743854226 +/- 0.3672980240022614  
y intercept 2.1208469582905676 +/- 0.28350968135605453

# adding a prior

Adding a prior on the parameters is equivalent to **regularization**.

Try to prevent overfitting and/or identify which parameters can be left out by "stiffening" the model.

## Ridge regression

$$\ln P(\boldsymbol{\theta}) = -\frac{1}{2} \left[ (\mathbf{y} - \mathbf{M}\boldsymbol{\theta})^T \mathbf{C}^{-1} (\mathbf{y} - \mathbf{M}\boldsymbol{\theta}) + \ln((2\pi)^N |\mathbf{C}|) + \lambda \boldsymbol{\theta}^T \boldsymbol{\theta} \right] \quad (331)$$

where  $\lambda$  is a free parameter that regulates the strength of the prior. And the maximum posterior solution will be

$$\hat{\boldsymbol{\theta}} = \left( \mathbf{M}^T \mathbf{C}^{-1} \mathbf{M} + \lambda \mathbf{I} \right)^{-1} \mathbf{M}^T \mathbf{C}^{-1} \mathbf{y} \quad (332)$$

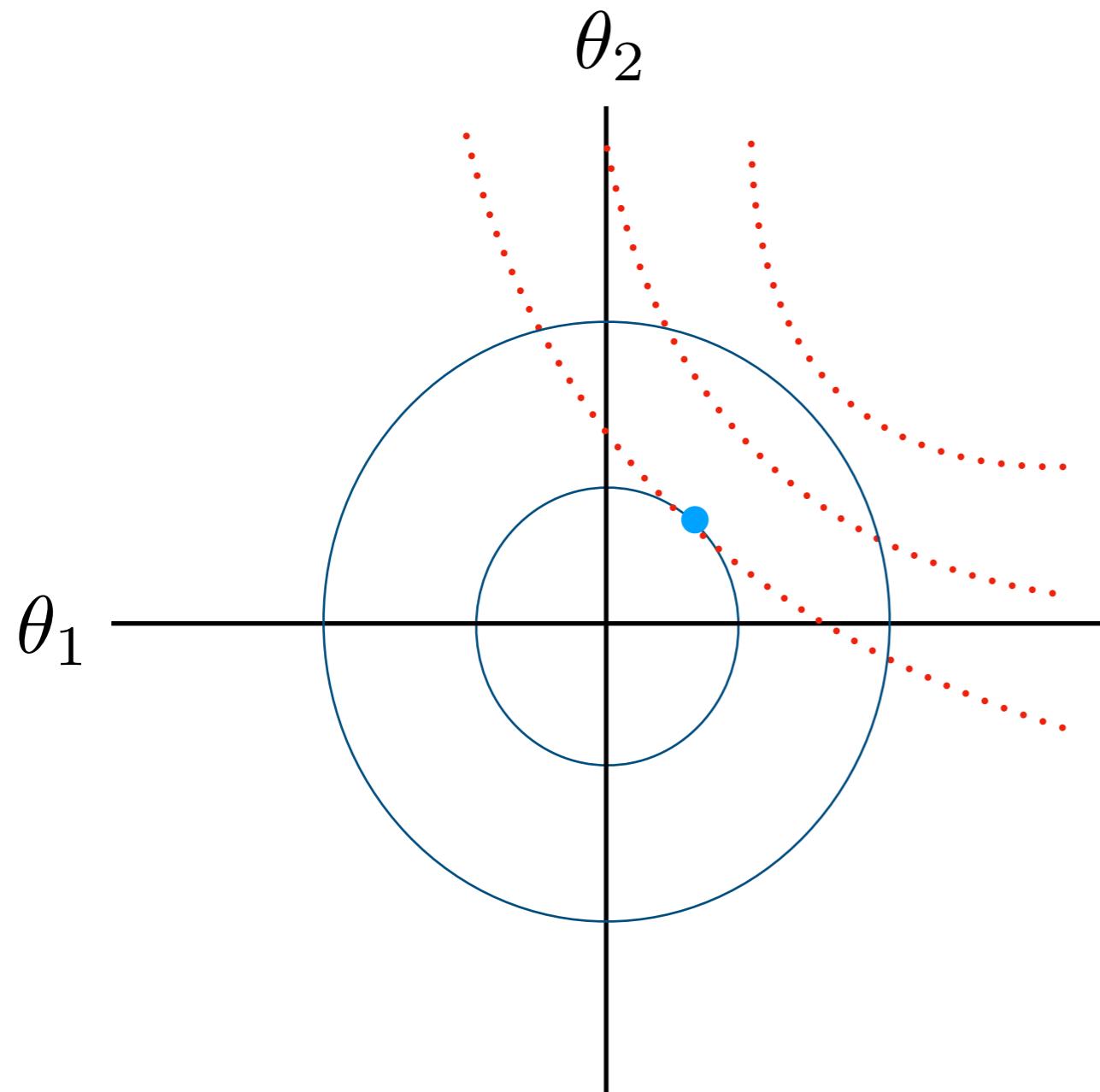
**LASSO regression** (least absolute shrinkage and selection operator).

$$\ln P(\boldsymbol{\theta}) = -\frac{1}{2} \left[ (\mathbf{y} - \mathbf{M}\boldsymbol{\theta})^T \mathbf{C}^{-1} (\mathbf{y} - \mathbf{M}\boldsymbol{\theta}) + \ln((2\pi)^N |\mathbf{C}|) + \lambda \|\boldsymbol{\theta}\|_1 \right] \quad (333)$$

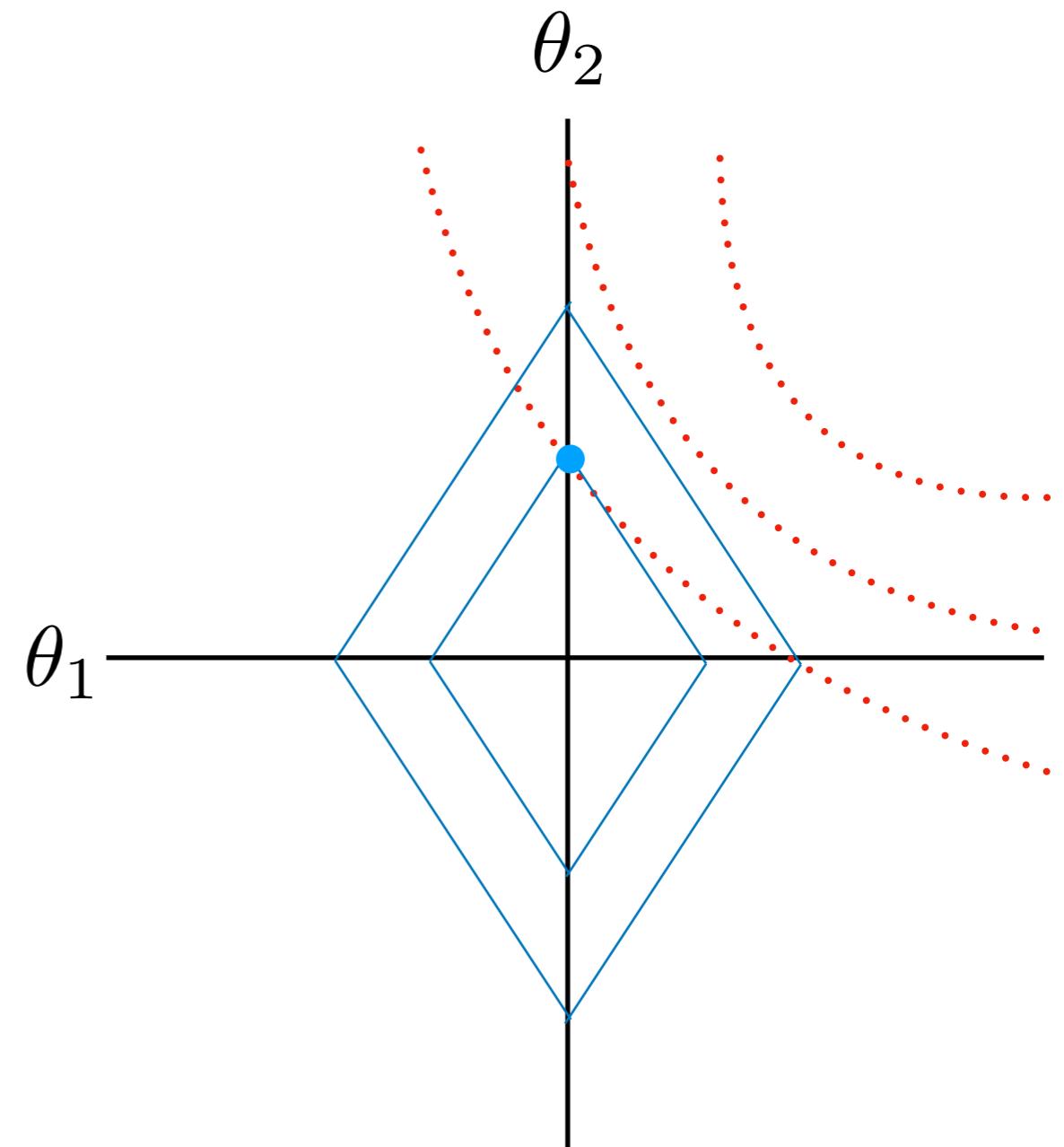
$$\|\boldsymbol{\theta}\|_1 = \sum_i |\theta_i|$$

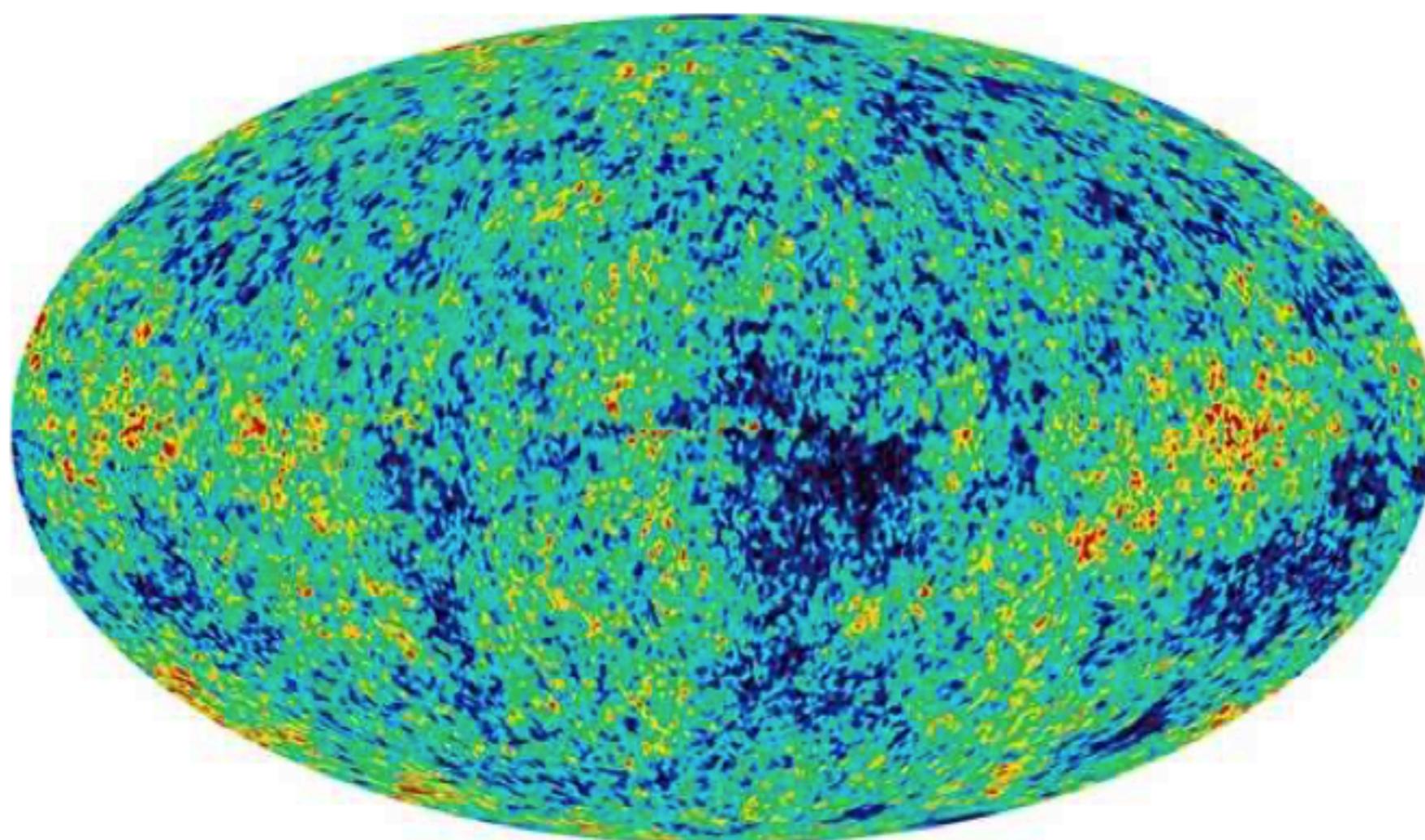
# Regularization or priors in regression

Ridge regression

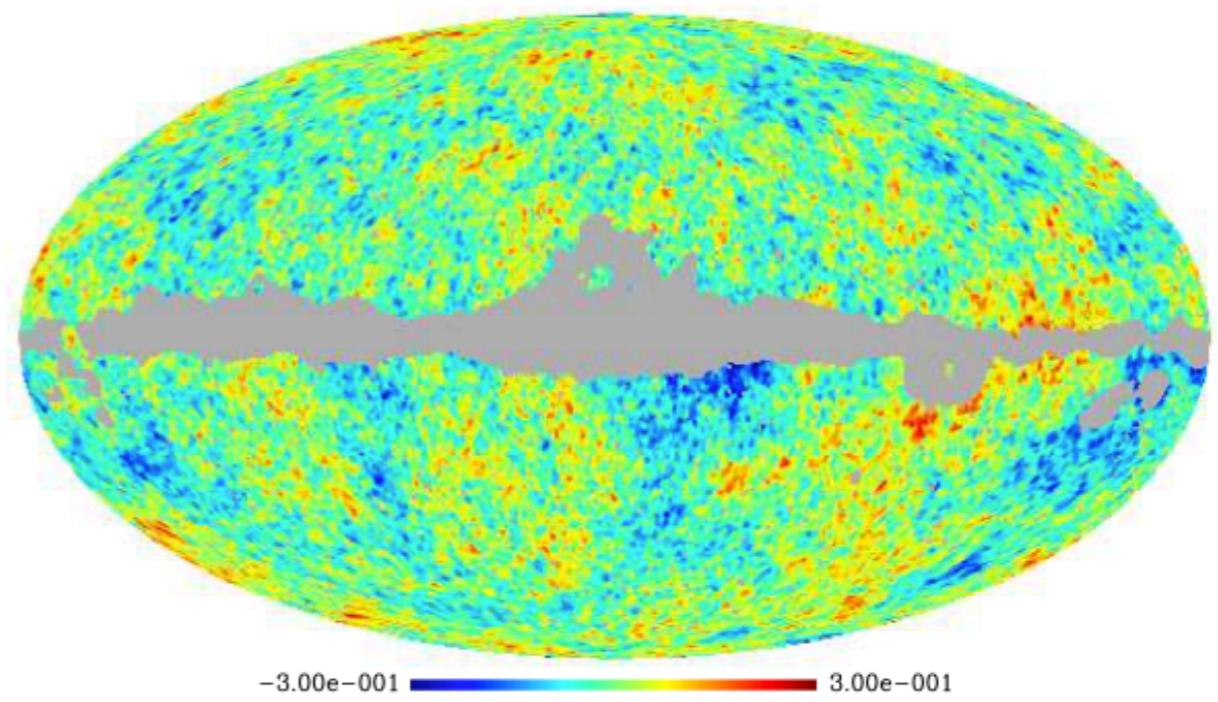


LASSO regression





**Wiener filter used to  
fill in masked regions**



# Robustness & breakdown point

**robustness** - the insensitivity of a statistic to contamination of the sample or to assumptions about distribution.

**breakdown point** - The fraction of the data that can be contaminated with data that is arbitrarily distributed before giving an arbitrarily wrong answer. The maximum breakdown point is 0.5 because at this point it would not be possible to differentiate between the contamination and the non-contamination.

The median has a breakdown point of 0.5. The mean has a breakdown point of 0.

The robustness can be increased by:

- trimming or culling - A fraction  $\alpha$  of the data that lies furthest from the model is removed and the statistic or model is refit to the remainder.
- M-estimators - Minimize

$$\sum_i \rho(d_i, \theta) \quad (432)$$

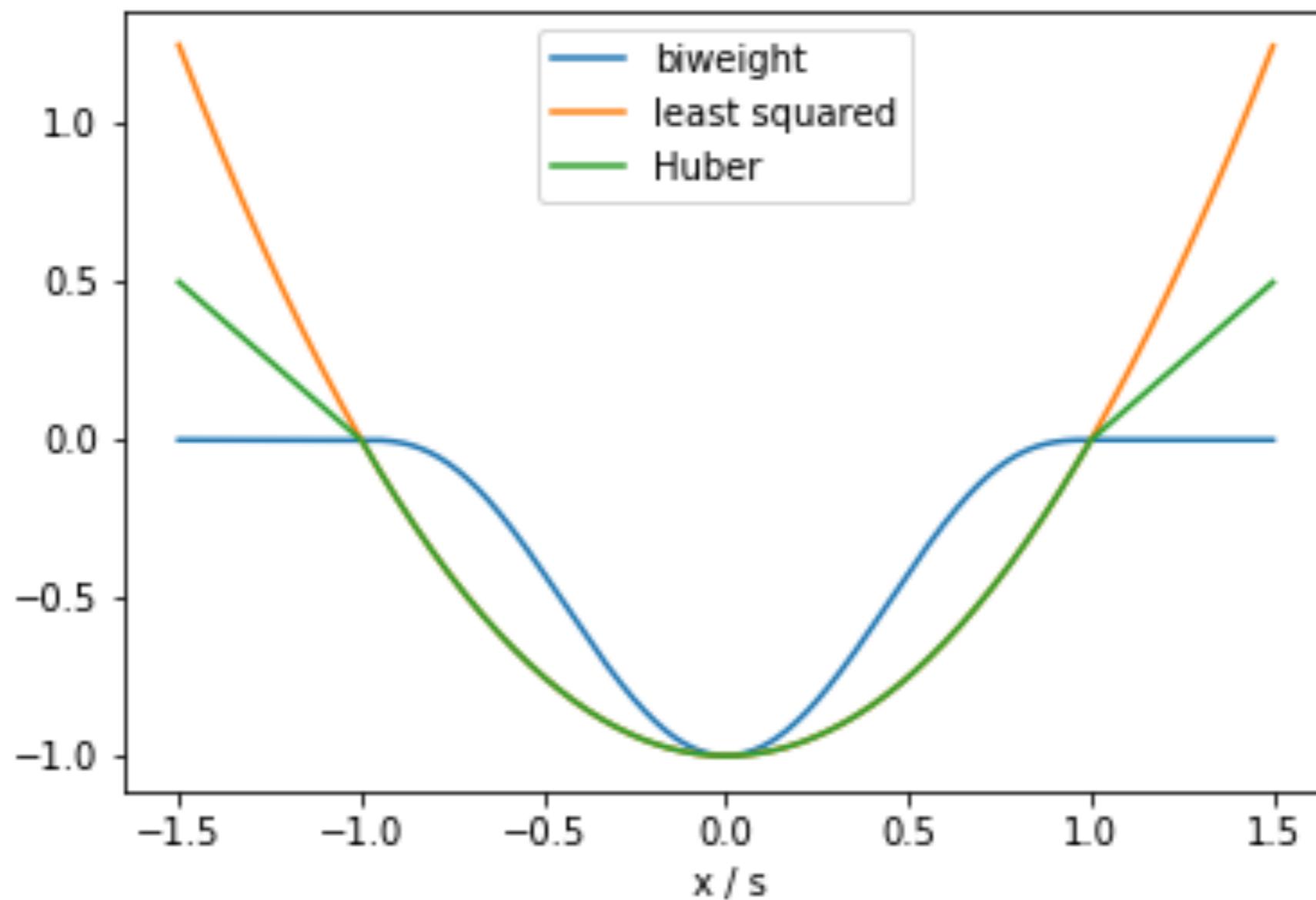
where  $\rho(d_i, \theta)$  is called the **loss function**.

- Least-squares -  $\rho(d_i, \theta) = [d_i - f_i(\theta)]^2$
- $\chi^2$  -  $\rho(d_i, \theta) = \frac{[d_i - f_i(\theta)]^2}{\sigma_i^2}$
- maximum likelihood (if factorable)-  $\rho(d_i, \theta) = \ln[\mathcal{L}(d_i|\theta)]$
- $L_1$  loss -  $\rho(d_i, \theta) = |d_i - f_i(\theta)|$
- Huber loss function -

$$\rho(x) = \begin{cases} \left(\frac{x}{s}\right)^2 & , \quad |x| < s \\ \frac{|x|}{s} & , \quad |x| > s \end{cases} \quad (433)$$

- Tukey's biweight

## M-estimator loss functions





# Hypothesis Testing

The basic steps in any hypothesis test are as follows:

- 1 State the hypothesis as a well posed true or false question. The goal is to falsify this question.
- 2 Choose or invent a statistic (called a **goodness of fit statistic**) that is affected by the truth of the hypothesis.
- 3 Calculate the value of the statistic with the data.
- 4 Determine by analytic or numerical methods the probability distribution of the statistic given that the hypothesis is true. Identify a direction or directions where the probability of getting that values for the statistic get less and less probable. This is usually as the statistic becomes very large absolutely or in magnitude.
- 5 With this distribution, calculate how probably it is for the statistic to be further in the direction of bad fits than the value calculated using the data.
- 6 If this probability is *sufficiently improbable* the hypothesis is ruled out. If it is not sufficiently improbable the hypothesis is *consistent* with this statistic.

Errors in hypothesis testing are by tradition categorized into two types:

- **Type I errors** - This is the case where the hypothesis is rejected, but is in fact true. You might call this a false positive.
- **Type II errors** - This is the case where the hypothesis is not rejected, but is in fact false. You might call this a false negative.

# Example: Test for difference of means

The statistic used is

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The null hypothesis is that the means of the populations are equal  $\mu_1 = \mu_2$ .  
This statistic is normally distributed.

# Example: Test for difference of means

## **Student's t test for the difference of two means.**

We might not know the measurement errors and need to estimate them from the data in which case the statistic

$$T = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

is used.  $s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$ . With the same null hypothesis this statistic has very nearly a t-distribution with

$$\nu = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

degrees of freedom. This is called the

## **DEMO**

### **t-test.py**

two data sets

**A = 97, 90, 95, 90, 101, 99, 99, 107, 102, 95**

**B = 101, 94, 93, 96, 94, 97, 94, 98, 98, 90, 90, 95**

Null Hypothesis 1 : The means are the same.

Null Hypothesis 2 : The the difference in the means is  $\Delta\mu$

# F test for the difference in the variance between two samples

We might also wonder if two populations have the same variance. You can test this with the statistic

$$f = \frac{S_1^2}{S_2^2}$$

This is of the form

$$\frac{X_\alpha^2/\alpha}{X_\beta^2/\beta}$$

where  $X_\alpha^2$  is a  $\chi_\alpha^2$  distributed variable. In this case  $\alpha = n_1 - 1$  and  $\beta = n_2 - 1$ . Such a ratio has a **F-distribution**, specifically  $F_{n_1-1, n_2-1}$ . The pdf is

$$p_F(f) = \frac{\alpha^{\alpha/2} \beta^{\beta/2} f^{\alpha/2-1}}{B(\alpha/2, \beta/2)(\alpha f + \beta)^{(\alpha+\beta)/2}}$$

where  $B(\alpha, \beta)$  is the beta function. Thus this is called the **F-test** for the difference of two variances.

**DEMO**

**f-test.py**

# $\chi^2$ test for the constancy of a signal

We (again) have  $n$  independent normally distributed data points,  $x_i$ .

**Null Hypothesis I** *The signal is constant, its value is equal to  $\mu$  and the errors are Gaussian distributed with the known variance  $\sigma$ .*

Here  $\mu$  is some fixed value that is not derived from the data, maybe zero. The measurement errors also fixed and known. The likelihood with this hypothesis is

$$p(x_i|\mu) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

statistic

$$\chi^2 = \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^2}$$

We know from section ?? that this statistic is  $\chi^2$  distributed with  $n$  degrees of freedom. We can calculate  $\chi^2$  with our data and find the cumulative probability up to this value  $F_{\chi_n^2}(X^2)$ . If this is large then we can say the a mean of  $\nu$  is ruled out at the  $1 - F_{\chi_n^2}(X^2)$  confidence level.

**Null Hypothesis II** *The signal is constant and the errors are Gaussian distributed with the known variance  $\sigma$ .*

statistic

$$\chi^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\sigma^2}$$

with the sample mean  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  in place of an hypothesized mean. This statistic will not be  $\chi_n^2$  distributed however.

Decomposing the data vector

$$\begin{aligned} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} &= \bar{x} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix} \\ &= (\mathbf{x} \cdot \hat{\mathbf{m}}) \hat{\mathbf{m}} + [\mathbf{x} - (\mathbf{x} \cdot \hat{\mathbf{m}}) \hat{\mathbf{m}}] \end{aligned}$$

$$\hat{\mathbf{m}} = (1, 1, \dots) / \sqrt{n}$$

This type of hypothesis is akin to doing Bayesian model selection in that it gives a criterion for rejecting a model irrespective of the specific values of the parameters.

**Null Hypothesis III** *Given that the signal is constant, its value is equal to  $\mu$  and the errors are Gaussian distributed with the known variance  $\sigma$ .*

A statistic based only on the first part of the decomposition (375), the part not included in the previous test.

$$\chi^2 = \sum_{i=1}^n \frac{(\bar{x} - \mu)^2}{\sigma^2} = \frac{n(\bar{x} - \mu)^2}{\sigma^2}$$

This is the magnitude of the projection of the data vector onto  $\hat{\mathbf{m}}$ , a one dimensional space. It has only one degree of freedom. and is thus  $\chi_1^2$  distributed.

- $\mathbf{M}$  takes a parameter vector to data space, but since these spaces have different dimensions it cannot cover all of data space.
- The matrix  $(\mathbf{M}^T \mathbf{C}^{-1} \mathbf{M})^{-1} \mathbf{M}^T \mathbf{C}^{-1}$  takes the data to the best fit model parameters - data space to parameters space.
- The matrix  $\mathbf{P} = \mathbf{M} (\mathbf{M}^T \mathbf{C}^{-1} \mathbf{M})^{-1} \mathbf{M}^T \mathbf{C}^{-1}$  projects the data onto the subspace that has direct affect on the parameters of the model. This is a  $k$  dimensional subspace because there are  $k$  parameters. The extra  $\mathbf{M}$  goes from parameter space to data space.
- The matrix  $\hat{\mathbf{P}} = \mathbf{I} - \mathbf{M} (\mathbf{M}^T \mathbf{C}^{-1} \mathbf{M})^{-1} \mathbf{M}^T \mathbf{C}^{-1}$  projects the data into the subspace that does not affect the parameters. These are the degrees of freedom that are not absorbed by fitting the model.

Any data vector can be decomposed into orthogonal parts

$$\mathbf{y} = \mathbf{P}\mathbf{y} + \bar{\mathbf{P}}\mathbf{y} \quad (386)$$

$$= \mathbf{y}_k + \mathbf{y}_{n-k} \quad (387)$$

where  $\mathbf{y}_k$  is in the  $k$  dimensional space that "influences" the model parameters and  $\mathbf{y}_{n-k}$  is in the  $n - k$  that is independent of the parameters.  $X^2(\mathbf{y}, \hat{\theta})$  contains only the  $\mathbf{y}_{n-k}$  component of the data.

The result of this is that if you find the best fit linear model and then calculate  $X^2(\mathbf{y}, \hat{\theta})$  with it you would expect this statistic to be  $\sim \chi^2_{n-k}$  if the model is correct.

You can calculate the p-value using this distribution. If  $X^2(\mathbf{y}, \hat{\theta})$  is too large you can rule out this model.

# The F-test for model selection

One linear model with  $k$  parameters and another model with  $m = k - r$  parameters.  
 We make the  $\chi^2$  statistics for each model and relate them with

$$\chi_M^2 = \chi_K^2 + (\chi_M^2 - \chi_K^2) \quad (388)$$

$$= \chi_K^2 + \Delta\chi^2 \quad (389)$$

It can be shown that  $\chi_K^2$  and  $\Delta\chi^2$  are statistically independent and that  $\chi_M^2 \sim \chi_{n-m}^2$ ,  $\chi_K^2 \sim \chi_{n-k}^2$  and  $\Delta\chi^2 \sim \chi_r^2$ .

Their ratio

$$f = \frac{\Delta\chi^2}{\chi_k^2} \left( \frac{n-k}{r} \right) \quad (390)$$

is a  $F_{r,n-k}$  distributed variable.

In particular if we add one parameter to the model we expect that

$$f = \frac{\Delta\chi^2}{\chi_k^2} (n - k) \quad (391)$$

will be  $F_{1,n-k}$ . If the measured value of  $f$  has a small chance of occurring according to this distribution we conclude that it is justified to add this parameter.

# frequentist confidence intervals

The likelihood:

$$X^2(x, \theta) = X^2(x, \hat{\theta}) + (\theta - \hat{\theta})^T M^T C^{-1} M (\theta - \hat{\theta}) \quad (392)$$

$$\mathcal{L} = \frac{1}{\sqrt{(2\pi)^2 |C|}} e^{-\frac{1}{2} X^2(x, \theta)} \quad (393)$$

You can see that  $X^2(x, \hat{\theta})$  was completely ignored in the Bayesian parameter estimate, while it is the only part that the frequentist hypothesis testing for the model was based on.

The boundaries of the confidence region (or interval in one dimension) in  $\theta$ -space are contours of equal likelihood i.e.

$$(\theta - \hat{\theta})^T M^T C^{-1} M (\theta - \hat{\theta}) = \text{constant} \quad (394)$$

The confidence level is taken from a  $\chi_k^2$  distribution.

Note that this is different from a Bayesian "**credibility region**" where the posterior is integrated within the boundaries of the region.

# *LIKELIHOOD RATIO TEST*

*Applicable to nested models.*

*Chi-squared hypothesis testing for parameter values is a special case*

*Also used for model selection or model building*

*More general than just for the likelihood Gaussian case*

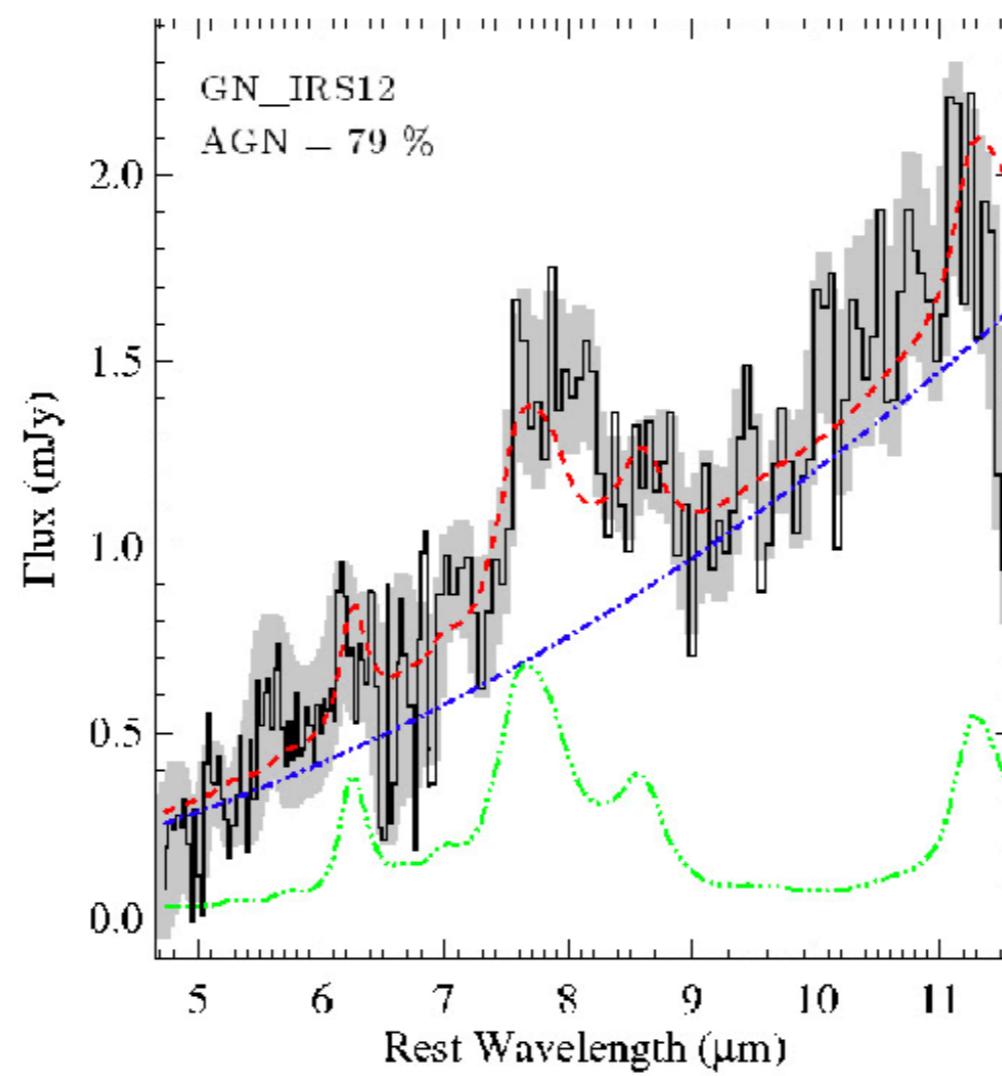
$H_0$     *The null hypothesis is that the simpler model is correct, which corresponds to the more complex model with some parameters fixed to specific values.*

$H_1$     *The alternative hypothesis is that the more complex model is required to explain the data*

$$\lambda \equiv -2 \ln \left[ \frac{\mathcal{L}(x|\hat{\theta}_o)}{\mathcal{L}(x|\hat{\theta})} \right]$$

$\hat{\theta}_o$     *The maximum likelihood parameters within the restricted, simpler, model.*

$\hat{\theta}$     *The maximum likelihood parameters within the unrestricted, more complex, model.*



## *LIKELIHOOD RATIO TEST*

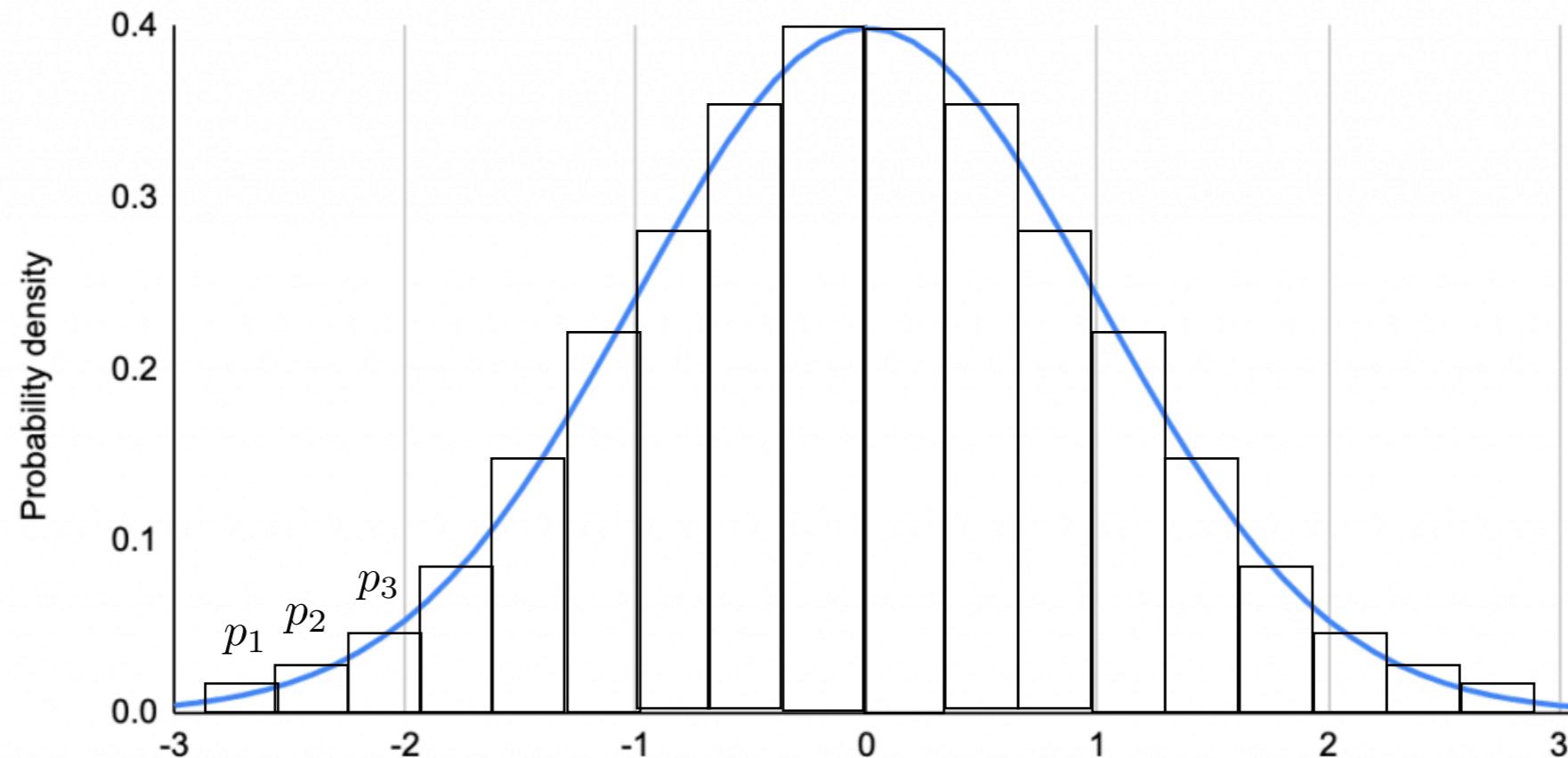
**Wilk's theorem** states that as the sample size goes to infinity,  $\lambda$  , will always approach  $\chi^2_{k-k_o}$  distributed as long as the  $\hat{\theta}_o$  does not lie on a boundary of the parameter space.

This is a justification for using the  $\lambda$  statistic as a general method for model selection when the amount of statistically independent data is large.



# Binned $\chi^2$ test

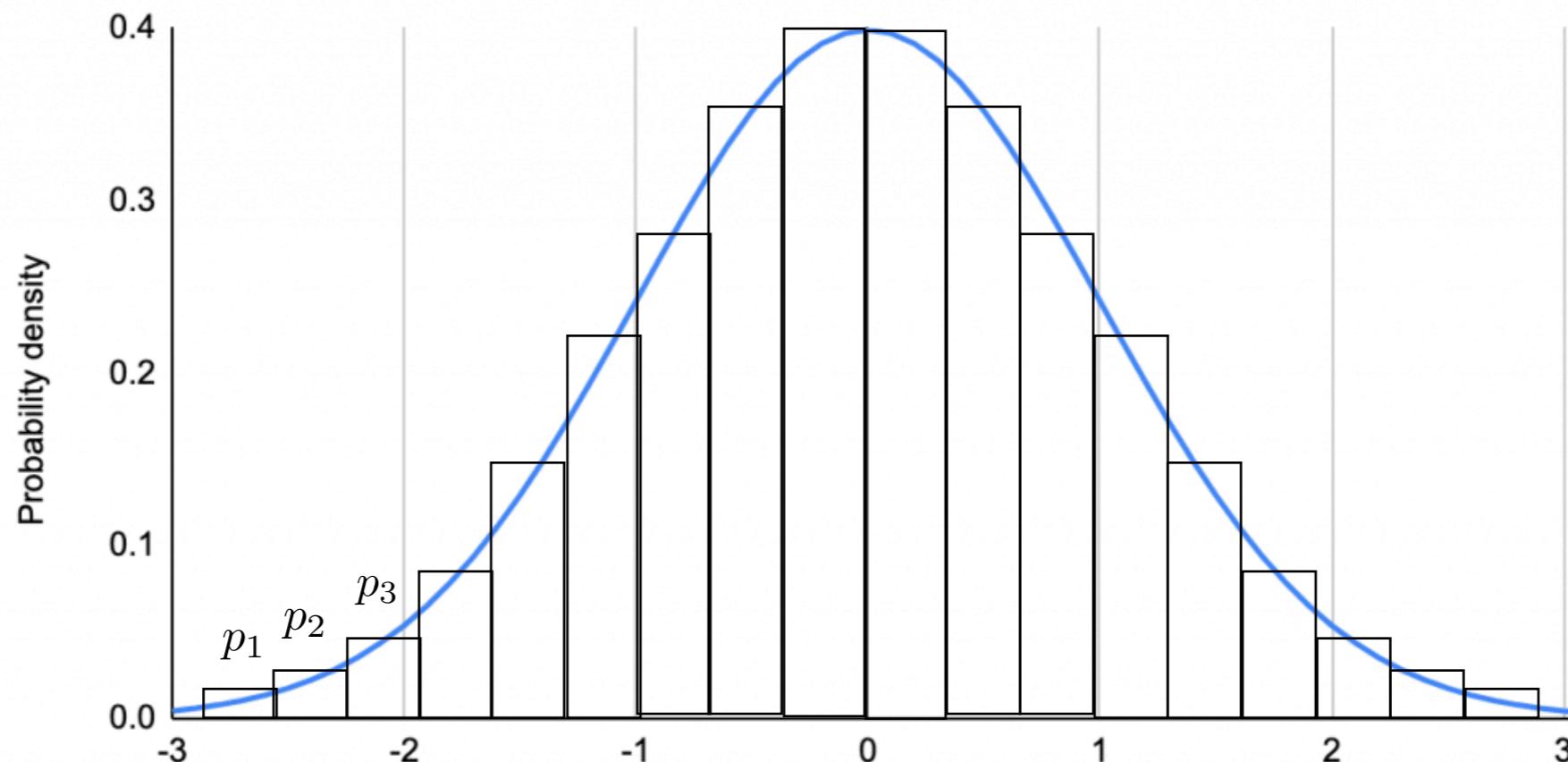
*Frequentist test for a distribution. For example mass function or luminosity function.*



# Binned $\chi^2$ test

## Multinomial Distribution

$$\begin{aligned} P(n_1, n_2, \dots, n_k | N, \{p_i\}) &= \frac{N!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} \\ &= \frac{N!}{\prod_{i=1}^k n_i!} \prod_{i=1}^k p_i^{n_i} \end{aligned}$$



# Binned $\chi^2$ test

$$P(\{n_i\}|N, \{p_i\}) = \frac{N!}{\prod_{i=1}^k n_i!} \prod_{i=1}^k p_i^{n_i}$$

The mean and variance of the number counts are

$$E[n_i] = Np_i \quad \text{Var}[n_i] = Np_i(1 - p_i)$$

We can expand the log probability around the average number counts using

$$\begin{aligned} \ln P(n_i = Np_i) &= N \ln(N) + \sum_i^k [Np_i \ln(p_i) - Np_i \ln(Np_i)] \\ &= 0 \qquad \qquad \qquad \sum_i p_i = 1 \end{aligned}$$

$$\left[ \frac{\partial}{\partial n_i} \ln P(\{n_i\}) \right]_{n_i=Np_i} = [\ln(p_i) - \ln(n_i)]_{n_i=Np_i} = -\ln N$$

$$\left[ \frac{\partial^2}{\partial n_i^2} \ln P(\{n_i\}) \right]_{n_i=Np_i} = \left[ -\frac{1}{n_i} \right]_{n_i=Np_i} = -\frac{1}{Np_i}$$

# Binned $\chi^2$ test

So the expansion is

$$\begin{aligned}\ln P(\{n_i\}) &\simeq -\ln N \sum_i (n_i - Np_i) - \sum_i \frac{1}{2Np_i} (n_i - Np_i)^2 + \mathcal{O}[(n_i - Np_i)^3] \\ &= -\sum_i \frac{1}{2Np_i} (n_i - Np_i)^2 + \mathcal{O}[(n_i - Np_i)^3] \quad \text{using } \sum_i n_i = N , \quad \sum_i p_i = 1\end{aligned}$$

The distribution of the counts in bins is approximately normal if the number counts in each bin is high

$$P(\{n_i\}) \simeq \frac{1}{\sqrt{(2\pi)^k N^k \prod_i^k p_i}} \exp \left[ -\sum_i^k \frac{(n_i - Np_i)^2}{Np_i} \right]$$

$k$  - number of bins

# Binned $\chi^2$ test

$$P(\{n_i\}) \simeq \frac{1}{\sqrt{(2\pi)^k N^k \prod_i^k p_i}} \exp \left[ - \sum_i^k \frac{(n_i - Np_i)^2}{Np_i} \right]$$

$k$  - number of bins

From this multivariate normal distribution we can form a chi-squared as before

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - Np_i)^2}{Np_i}$$

will be approximately  $\chi^2_{k-1}$  distributed because the one constraint that  $N = \sum_i$  all  $n_i$ 's are large.

# Kolmogorov-Smirnov (KS) test

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \Theta(x_i \leq x)$$

$\hat{F}(x)$  is an unbiased estimator of the cumulative distribution,  $F(x)$ . The KS statistic is

$$D_n = \max_x |\hat{F}_n(x) - F(x)|$$

this statistic is

$$\lim_{n \rightarrow \infty} P(\sqrt{n}D_n \leq t) = H(t) = 1 - 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 t^2}$$

for large  $n$ . This is independent of the distribution that is being tested  $F(x)$ .

# two sample KS test

The hypothesis here is that both data sets come from the same data distribution. Let's say the empirical cumulative distributions for these two samples are  $\hat{F}_n(x)$  and  $\hat{G}_m(x)$ . The statistic is the maximum vertical distance between the two sample cumulative distributions

$$D_{mn} = \max |\hat{F}_n(x) - \hat{G}_m(x)|$$

This statistic is distributed like

$$\lim_{n \rightarrow \infty} P \left( \sqrt{\frac{mn}{m+n}} D_{mn} \leq t \right) = H(t)$$

for large  $n$ .

## Cremér-von Mises test

Like the KS test this test seeks to test the consistency of data with a given distribution, but uses the statistic

$$\begin{aligned} T_{CM} &= n \int_{-\infty}^{\infty} dF(x) [\hat{F}(x) - F(x)]^2 \\ &= \frac{1}{12n} + \sum_{i=1}^n \left( \frac{2i-1}{2n} - F(x_i) \right)^2 \end{aligned}$$

where  $x_1, \dots, x_n$  are the sorted data points.

## Anderson-Darling test

$$\begin{aligned} A_{AD}^2 &= n \int_{-\infty}^{\infty} dF(x) \frac{[\hat{F}_n(x) - F(x)]^2}{F(x)[1 - F(x)]} \\ &= -n - \sum_i^n \frac{2i-1}{n} [\ln(F(x_i)) - \ln(1 - F(x_{n+1-i}))] \end{aligned}$$

# *Numerical Goodness-of-fit*

***Problem:*** *If the distribution of the goodness-of-fit statistic is not known how can you do hypothesis test?*

*Often the data can be simulated given the null hypothesis.*

*If we simulate a large number of data sets we can calculate the statistic for each one and find its distribution imperially.*

*The p-value can be found by making an imperial cumulative distribution function and comparing it to the statistic calculated from the real data.*

# *Numerical Goodness-of-fit*

***Problem 2:*** *Do a simulation of the data usually requires adopting some particular parameter values.*

*How can we do a global goodness-of-fit test if our test is based on specific parameter values?*

# *Numerical Goodness-of-fit*

*Solution:*

- 1) Make a fake data set  $D^*$  with the maximum likelihood model  $\hat{\beta}$**
- 2) Find the best-fit parameters for this data set  $\beta^*$**
- 3) Calculate the goodness-of-fit statistic  $A^* = A(D^*, \beta^*)$**
- 4) Repeat 1-3**
- 5) Compare the cumulative distribution of  $A^*$  to the observed  $A$  to find its significance.**

*The distribution of  $A^*$  will converge to the distribution of  $A(D, \hat{\beta})$  as the number of simulation goes to infinity.*

# rank statistics

If the data is sorted from least value to largest value the **rank** of a data point is where it appears in this list. In other words if a data point  $x_i$  has a rank  $X_i$  there are  $X_i - 1$  data points with smaller values.

The advantage to using rank instead of value is that we know the distribution of the ranks no matter what the distribution of the values is. It is uniformly distributed between 1 and  $n$ .

Are two variables correlated?

**Pearson's correlation coefficient** (not a rank statistic)

$$r_{xy} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2} \sqrt{\sum_i^n (y_i - \bar{y})^2}}$$

The statistic

$$t = r_{xy} \sqrt{\frac{n - 2}{1 - r_{xy}^2}}$$

is t-distributed with  $n - 2$  degrees of freedom if the  $x_i$ 's and  $y_i$ 's are independent and normally distributed.

But what if they aren't normally distributed?

If ranks are  $x_i$  are  $X_i = \{1, 2, \dots, n\}$ .

$Y_i = \{1, 2, \dots, n\}$  - perfectly correlated

$Y_i = \{n, n-1, \dots, 1\}$  - perfectly anti-correlated

$$\bar{X} = \sum_{i=1}^n X_i = \sum_{i=1}^n i = \frac{1}{2}n(n+1)$$

and

$$\sum_{i=1}^n X_i^2 = \sum_{i=1}^n i^2 = \frac{1}{6}n(n+1)(2n+1)$$

The variance of both  $X_i$  and  $Y_i$  are

$$V_X = V_Y = \frac{1}{n} \sum_i (X_i - \bar{X})^2 = \frac{(n^2 - 1)}{12}$$

# Spearman's correlation coefficient

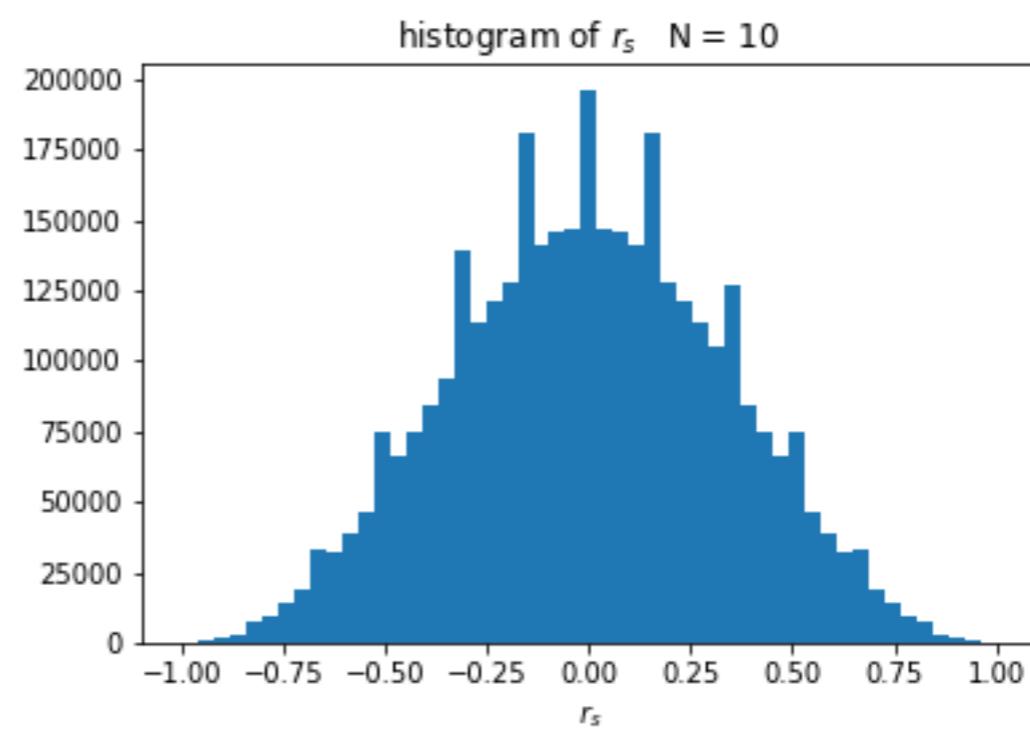
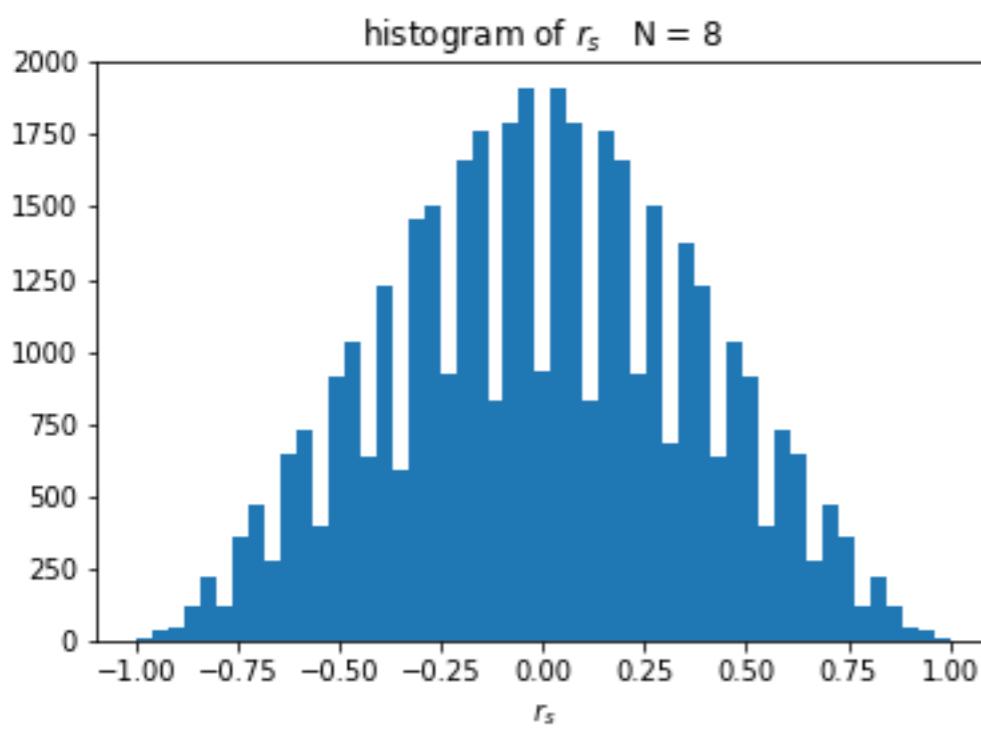
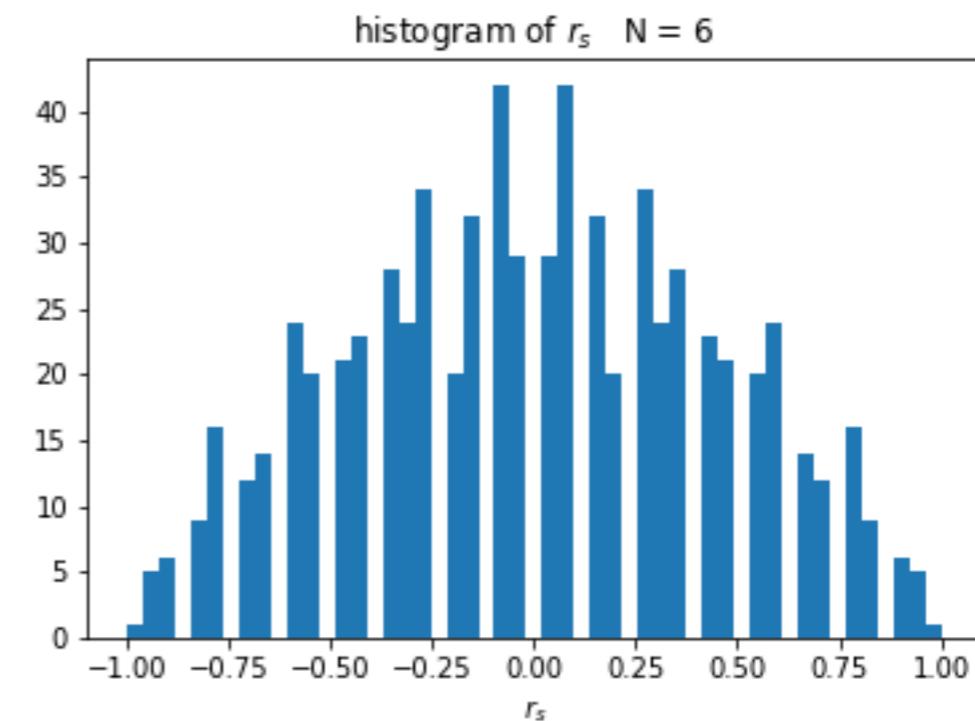
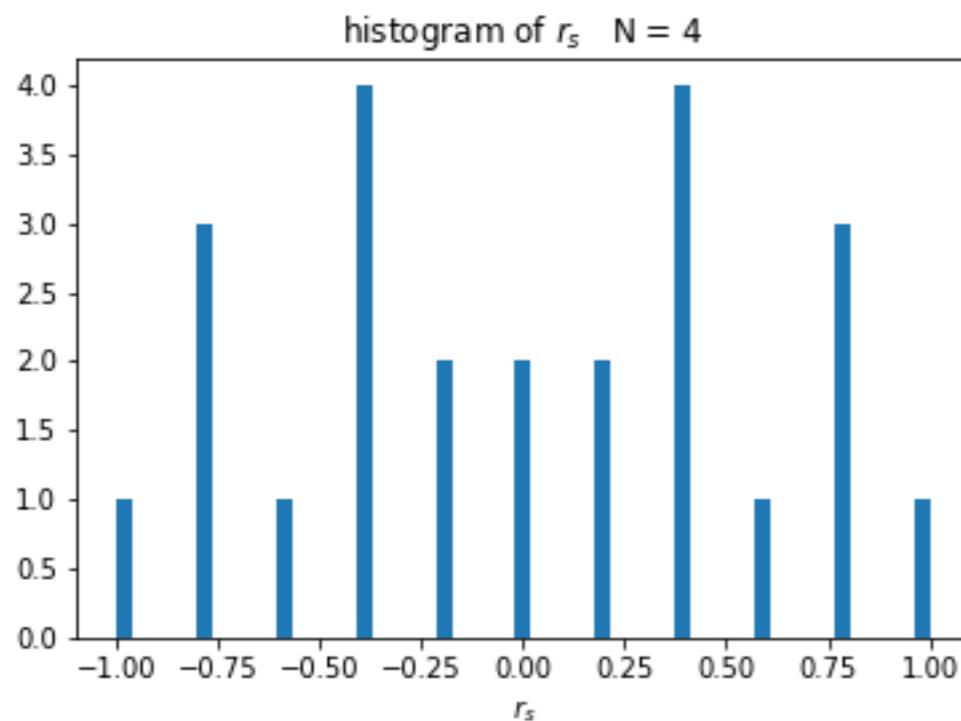
Spearman's correlation coefficient is the same thing as Pearson's, but using the ranks,  $X_i$  instead of the values  $x_i$ ,

$$r_s = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i^n (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}}$$

Also equal to

$$r_s = 1 - \frac{6}{n(n^2 - 1)} \sum_i (X_i - Y_i)^2$$

# Permutation test for Spearman's Correlation Coefficient



# Kendall's correlation coefficient

Let  $Q$  be the number of pairs of  $Y_i$ 's that are out of order, the number of inversions.  
Or

$$h_{ij} = \begin{cases} 1 & Y_i > Y_j \\ 0 & \text{otherwise} \end{cases} \quad Q = \sum_{i < j} h_{ij}$$

So for  $Y_i = \{1, 9, 6, 7, 5\}$   $Q = 5$ . Kendall's correlation coefficient is

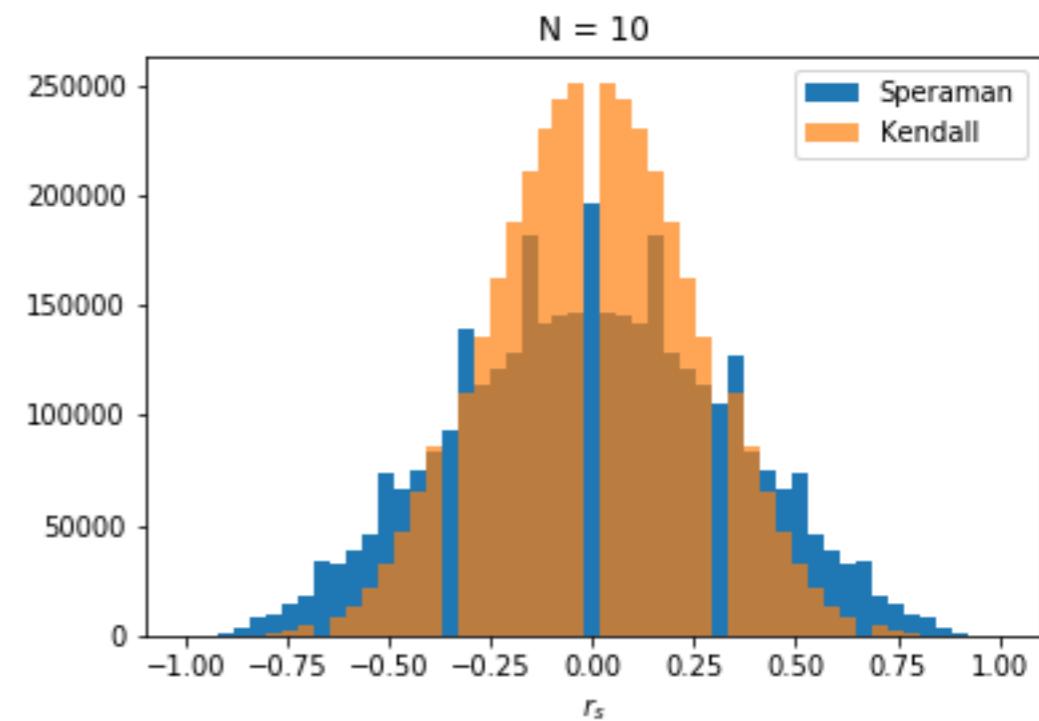
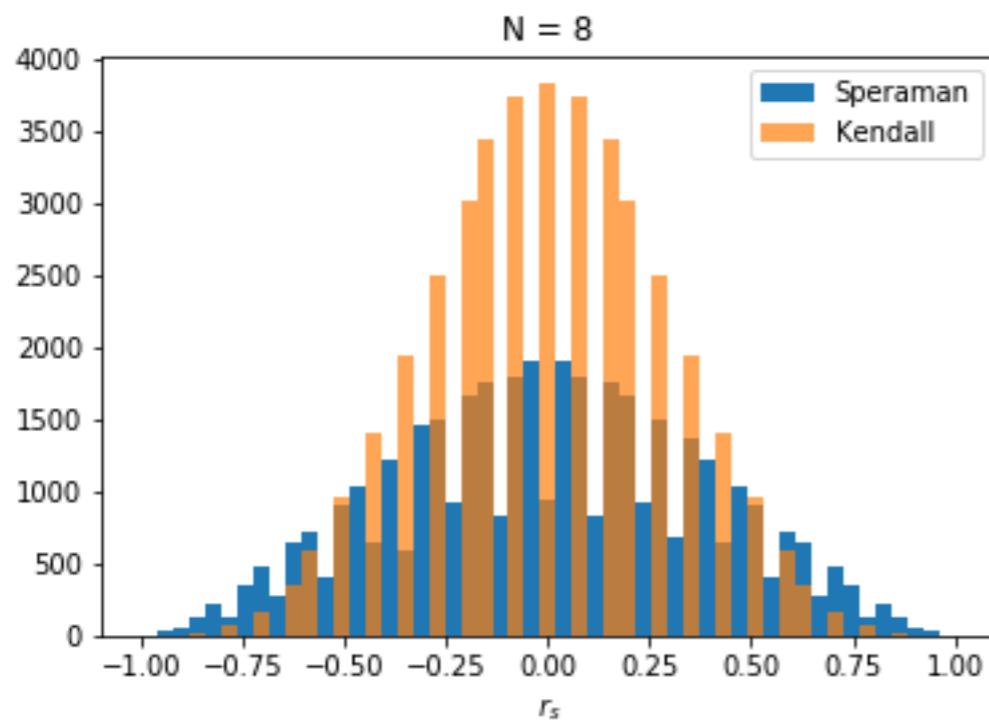
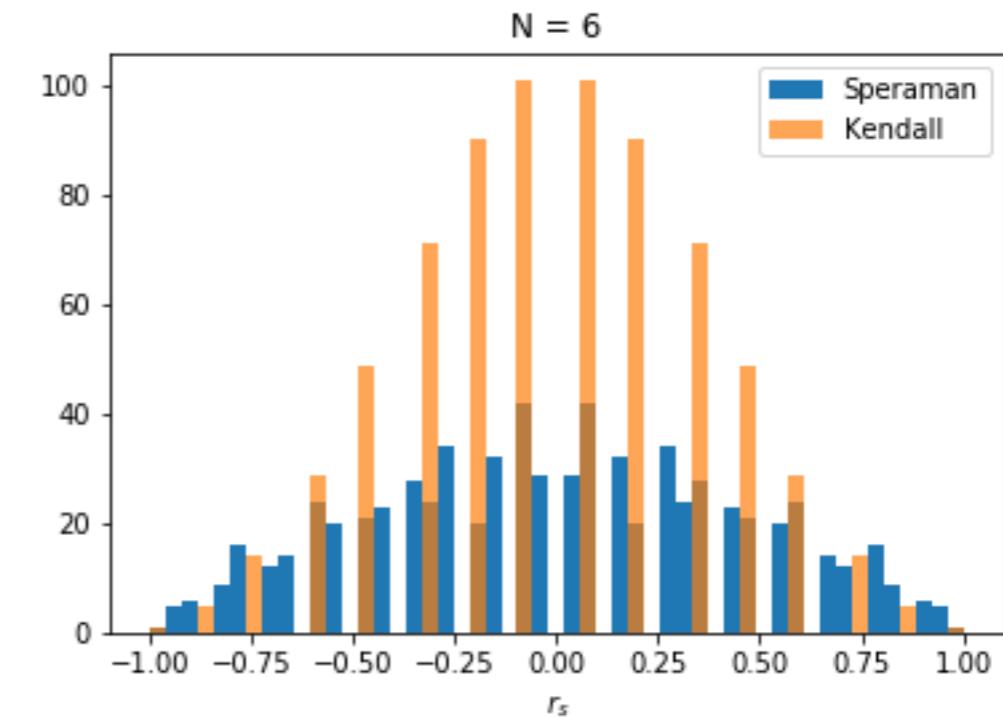
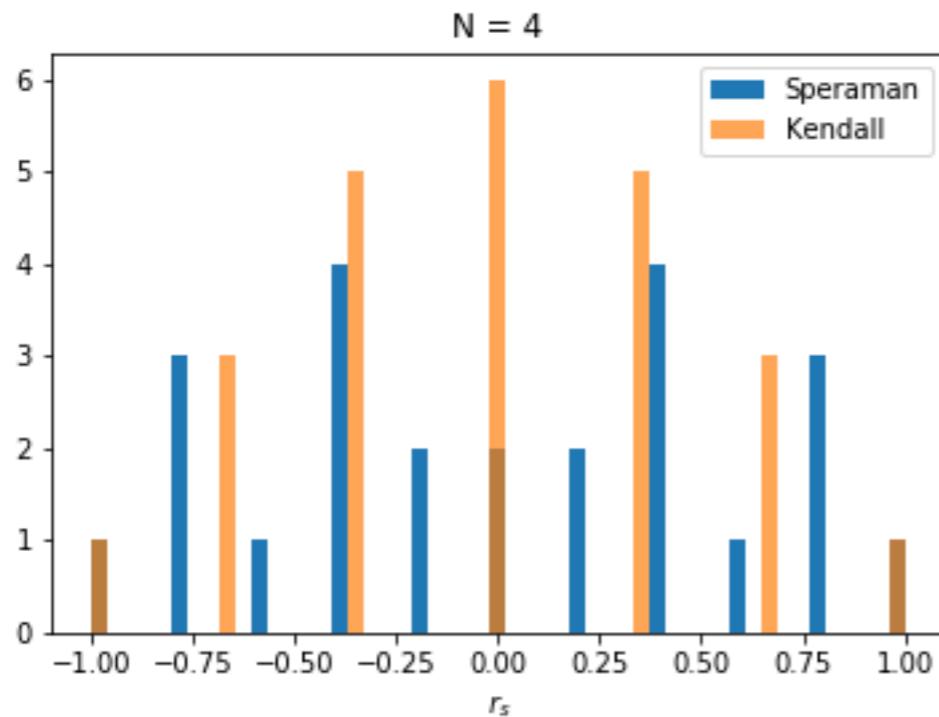
$$t = 1 - \frac{4Q}{n(n-1)}$$

$$t = \begin{cases} 1 & \text{perfect correlation} \\ 0 & \text{no correlation} \\ -1 & \text{perfect anticorrelation} \end{cases}$$

For  $n \gtrsim 10$

$$t \sim \mathcal{N} \left( 0, \sigma_t^2 = \frac{2(2n+5)}{9n(n-1)} \right)$$

# Permutation test for Spearman and Kendall's Correlation Coefficients



Wilcoxon's U (or the **Mann-Whitney test** or **rank-sum test**) is a rank test for the equality of the means of two samples.

$X_i$  are the ranks of one sample within the combined sample  $X$  and  $Y$ .

$$U = \sum_i X_i - \frac{1}{2} n_x(n_x + 1)$$

$U$  is always between 0 and  $n_x n_y$ .

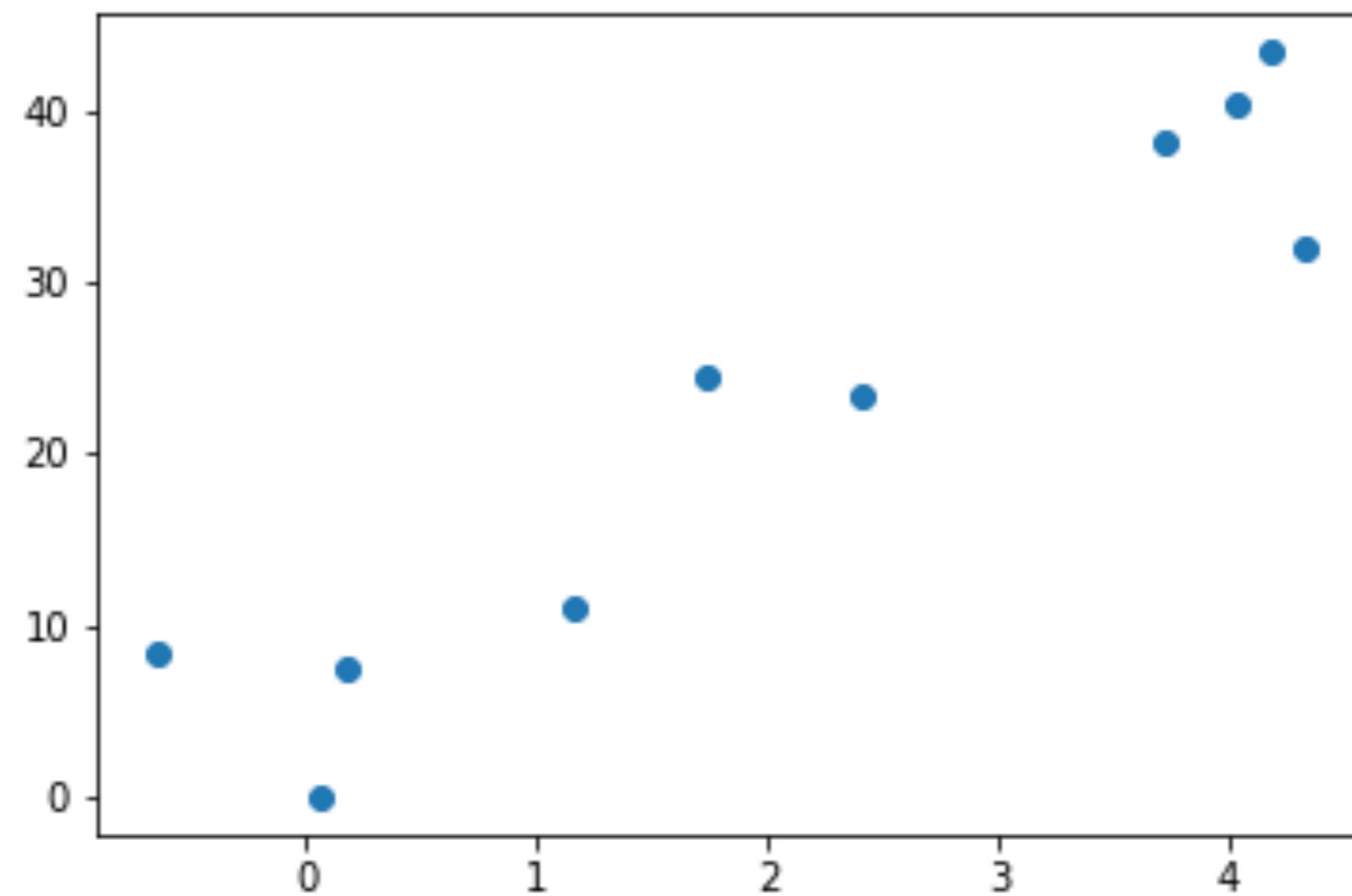
For moderate  $n$  (only  $n_x, n_y \gtrsim 8$ ) it is close to normally distributed with a mean and variance

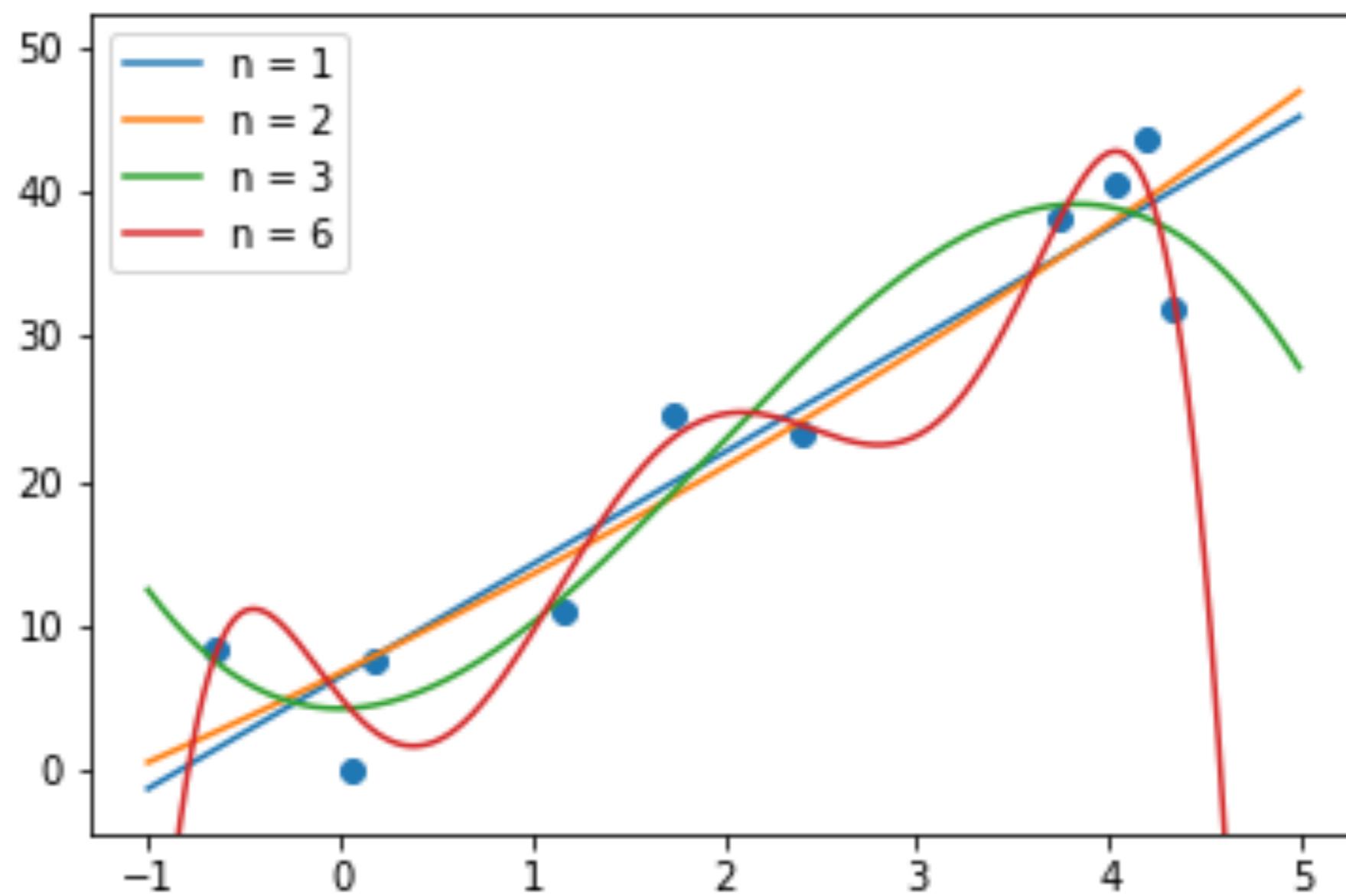
$$\langle U \rangle = \frac{n_x n_y}{2}$$

$$\sigma_U^2 = \frac{1}{12} n_x n_y (n_x + n_y + 1)$$

Under the hypothesis that the means of the samples are the same.







# Model Selection / Testing

## K-fold validation

independent of distribution  
at least in large N limit

- need a large amount of labeled data

$$\chi^2_{min}(\boldsymbol{x})$$

require that the two sided p-value is large

- only valid for a Linear Gaussian Model (LGM)
- not very sensitive

## likelihood ratio test

based of p-value of  $\lambda(\boldsymbol{x})$  statistic for a.  $\chi^2$  distribution

- only valid for **nested** LGMs
- could be calculated for non LGM by Monte Carlo

# Bayesian Model Selection

Posterior for model  $M_i$  using Bayes' theorem as in the parameter estimation case

$$P(M_i|\mathbf{D}) = \frac{P(\mathbf{D}|M_i)P(M_i)}{P(\mathbf{D})} = \frac{P(\mathbf{D}|M_i)P(M_i)}{\sum_i P(\mathbf{D}|M_i)P(M_i)}.$$

The **odds** of model 1 relative to model 2 is

$$O_{1,2} = \frac{P(M_1|\mathbf{D})}{P(M_2|\mathbf{D})} = \frac{P(\mathbf{D}|M_1)}{P(\mathbf{D}|M_2)} \frac{P(M_1)}{P(M_2)} = B_{1,2} \frac{P(M_1)}{P(M_2)}.$$

$B_{1,2}$ , the ratio of the model likelihoods, is called **Bayes's factor**.

We can write Bayes' theorem again with the model conditionality explicitly shown

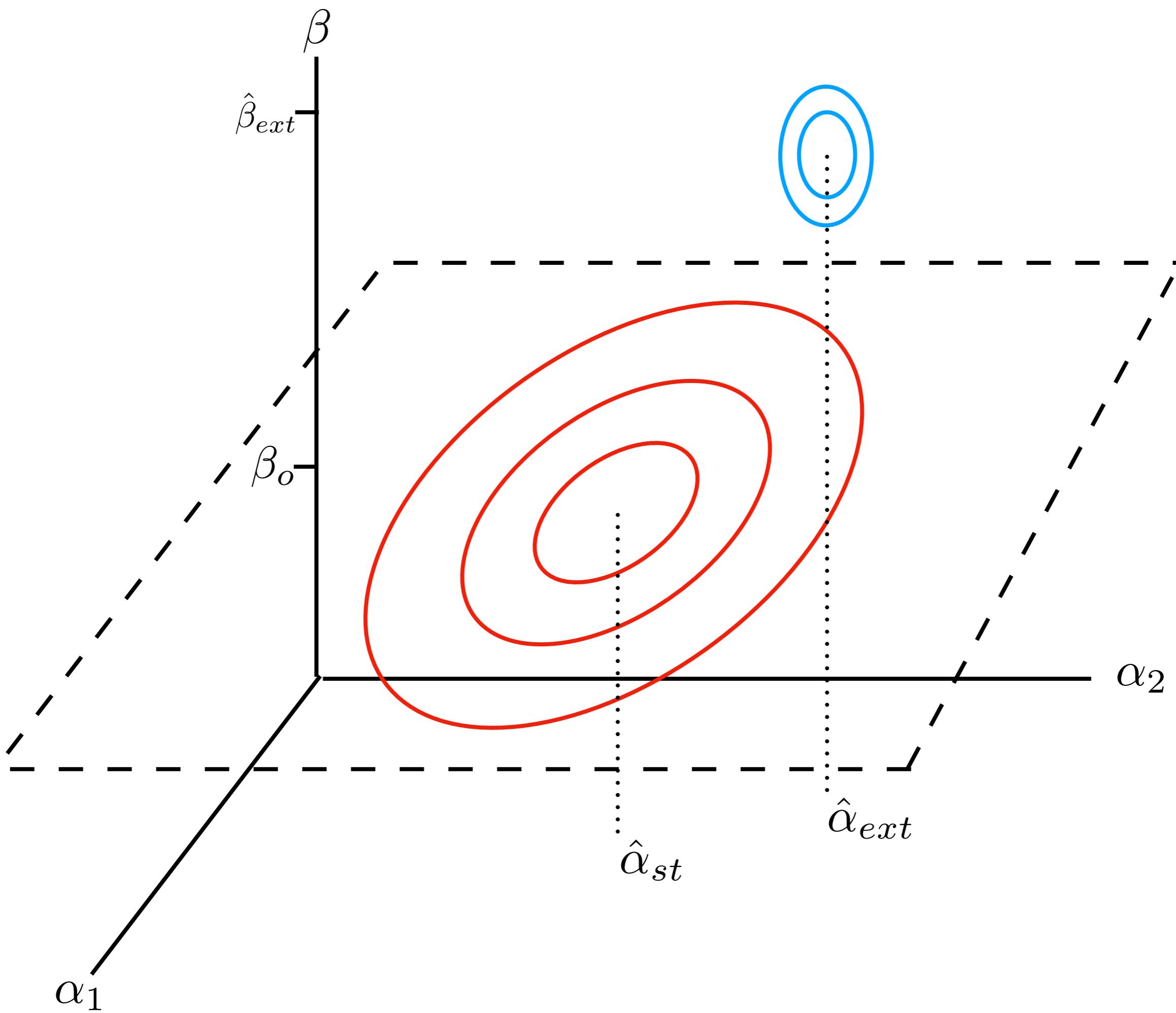
$$P(\boldsymbol{\theta}|\mathbf{D}, M) = \frac{P(\mathbf{D}|\boldsymbol{\theta}, M)P(\boldsymbol{\theta}|M)}{P(\mathbf{D}|M)}.$$

We can now see that  $P(\mathbf{D}|M)$  is actually the evidence:

$$P(\mathbf{D}|M_i) = \int_{-\infty}^{\infty} d\boldsymbol{\theta} P(\mathbf{D}|\boldsymbol{\theta}, M_i)P(\boldsymbol{\theta}|M_i) = \mathcal{E}(\mathbf{D}|M_i)$$

$$B_{1,2} = \frac{\mathcal{E}_1(\mathbf{D})}{\mathcal{E}_2(\mathbf{D})}$$

**Bayes' factor** is the ratio of the evidences calculated with the two models.



# Occam's factor

Does the data justify adding more parameters?

For **nested models** there will always be a parameter set where the likelihood is larger.

$$\mathcal{L}_{ex}(\hat{\boldsymbol{\theta}}_{ex}, \hat{\beta}) \geq \mathcal{L}_{st}(\hat{\boldsymbol{\theta}}_{st})$$

$$B_{2,1} = \frac{\int d\boldsymbol{\theta} \int d\beta \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}, \beta) \pi(\boldsymbol{\theta}, \beta)}{\int d\boldsymbol{\theta} \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}, \beta_o) \pi(\boldsymbol{\theta})} = \frac{\langle \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}, \beta) \rangle_{\boldsymbol{\theta}, \beta}^{\pi}}{\langle \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}, \beta_o) \rangle_{\boldsymbol{\theta}}^{\pi}}$$

If  $\pi(\boldsymbol{\theta}, \beta) = \pi(\boldsymbol{\theta})\pi(\beta)$

$$B_{2,1} = \frac{\langle \langle \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}, \beta) \rangle_{\boldsymbol{\theta}}^{\pi} \rangle_{\beta}^{\pi}}{\langle \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}, \beta_o) \rangle_{\boldsymbol{\theta}}^{\pi}}$$

$$\mathcal{V}_{\boldsymbol{\theta}}^{\mathcal{L}} \equiv \frac{1}{\mathcal{L}(\mathbf{D}|\hat{\boldsymbol{\theta}})} \int d\boldsymbol{\theta} \ \mathcal{L}(\mathbf{D}|\boldsymbol{\theta})$$

characteristic volumes to which the likelihood and prior contain the parameters

$$\mathcal{V}_{\boldsymbol{\theta}}^{\pi} \equiv \frac{1}{\pi(\hat{\boldsymbol{\theta}})} \int d\boldsymbol{\theta} \ \pi(\boldsymbol{\theta}) = \frac{1}{\pi(\hat{\boldsymbol{\theta}})}$$

An average of the likelihood over the prior can then be represented as

$$\langle \mathcal{L}(\boldsymbol{\theta}, \beta) \rangle_{\beta}^{\pi} \sim \left[ \frac{\mathcal{V}_{\beta}^{\mathcal{L}}}{\mathcal{V}_{\beta}^{\pi}} \right] \mathcal{L}(\boldsymbol{\theta}, \hat{\beta})$$

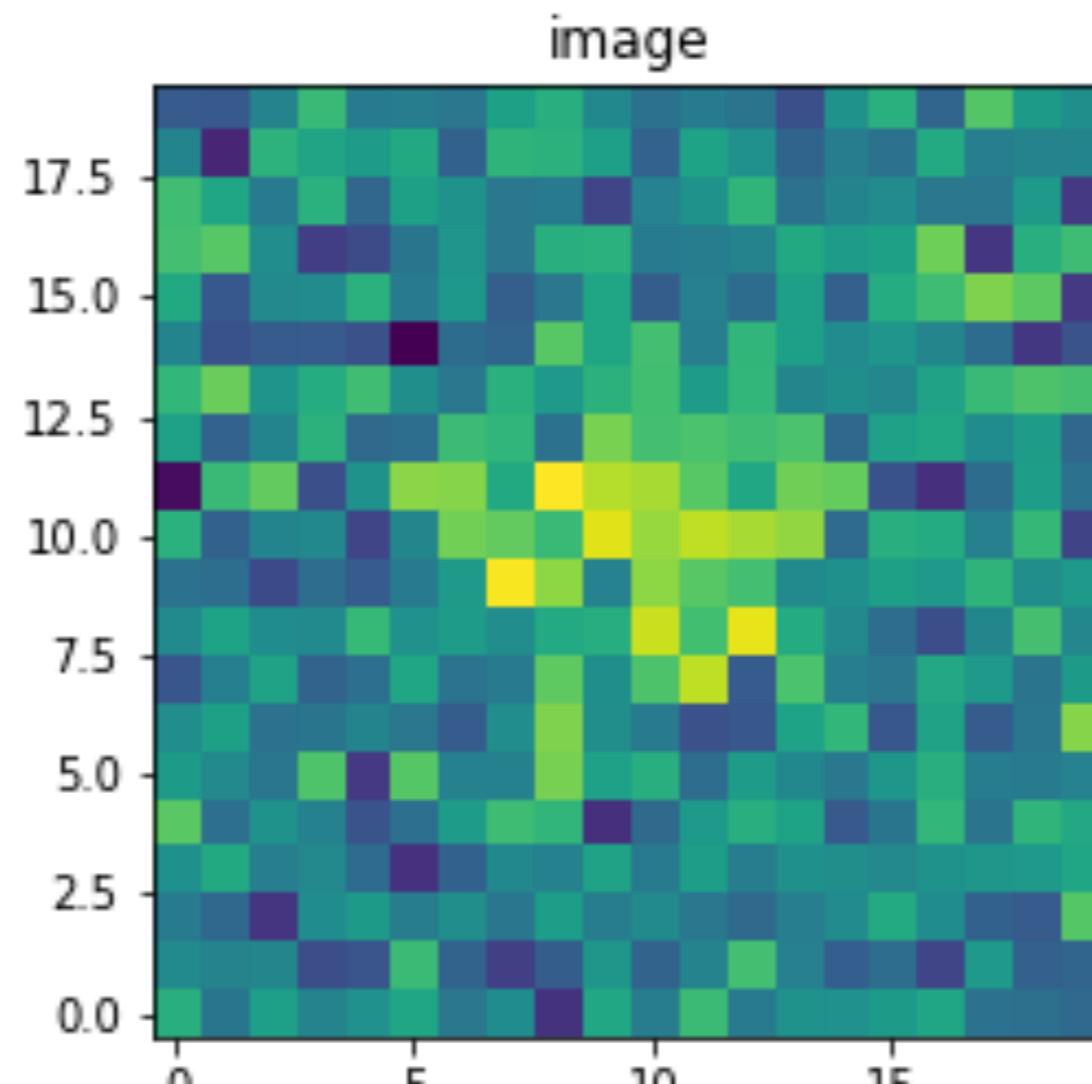
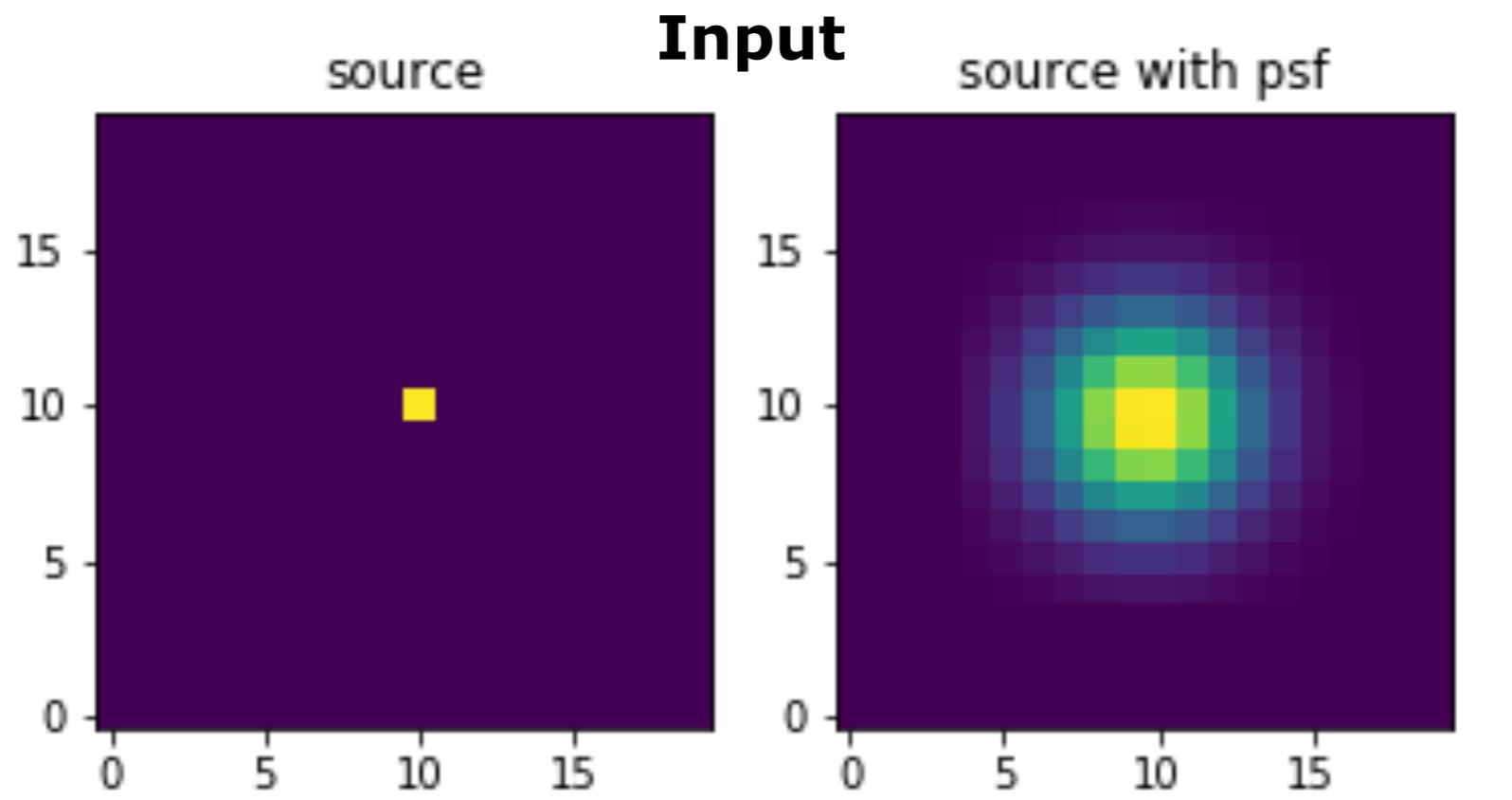
and Bayes' factor

$$B_{2,1} \sim \left[ \frac{\mathcal{V}_{\beta}^{\mathcal{L}}}{\mathcal{V}_{\beta}^{\pi}} \right] \frac{\left\langle \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}, \hat{\beta}) \right\rangle_{\boldsymbol{\theta}}^{\pi}}{\left\langle \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}, \beta_o) \right\rangle_{\boldsymbol{\theta}}^{\pi}}$$

An illustrative extreme case is one where the prior on  $\beta$  is very narrow compared to its constraint from the likelihood.

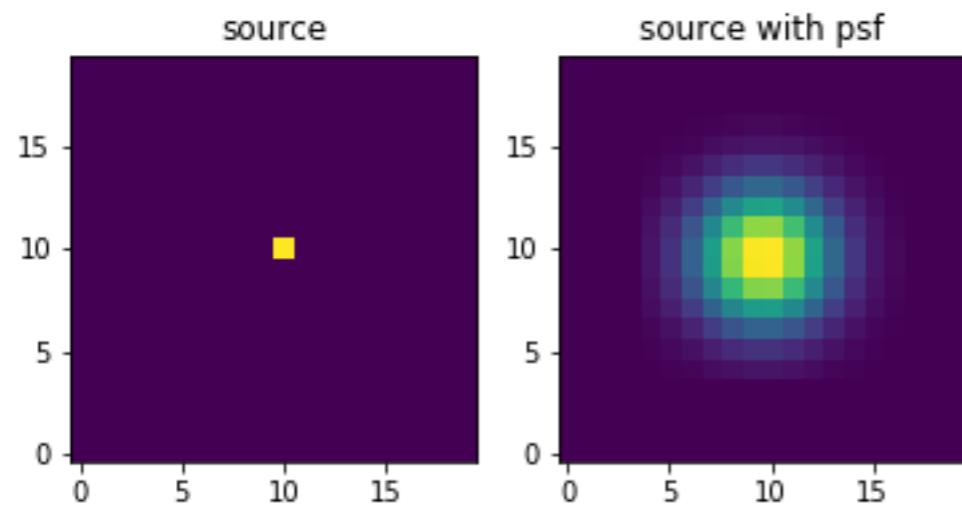
$$\begin{aligned}
 B_{2,1} &= \frac{\int d\theta \int d\beta \mathcal{L}(D|\theta, \beta) \pi(\theta) \pi(\beta)}{\int d\theta \mathcal{L}(D|\theta, \beta_o) \pi(\theta)} \\
 &\simeq \frac{\int d\theta \mathcal{L}(D|\theta, \beta_o) \pi(\theta) \int d\beta \pi(\beta)}{\int d\theta \mathcal{L}(D|\theta, \beta_o) \pi(\theta)} \\
 &\simeq \frac{\int d\theta \mathcal{L}(D|\theta, \beta_o) \pi(\theta)}{\int d\theta \mathcal{L}(D|\theta, \beta_o) \pi(\theta)} \\
 &\simeq 1
 \end{aligned}$$

No preference if the likelihood is not a function of the new parameter or the prior already constrains it much more strongly than the likelihood.

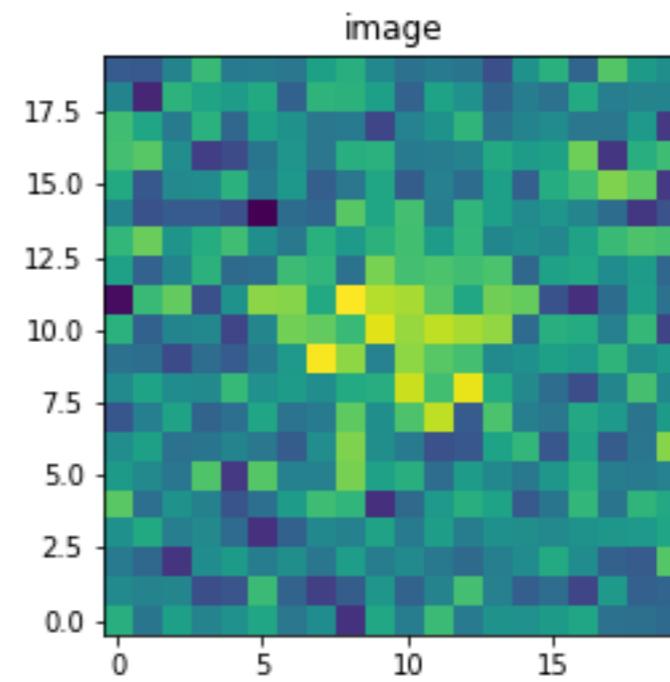


**Data**

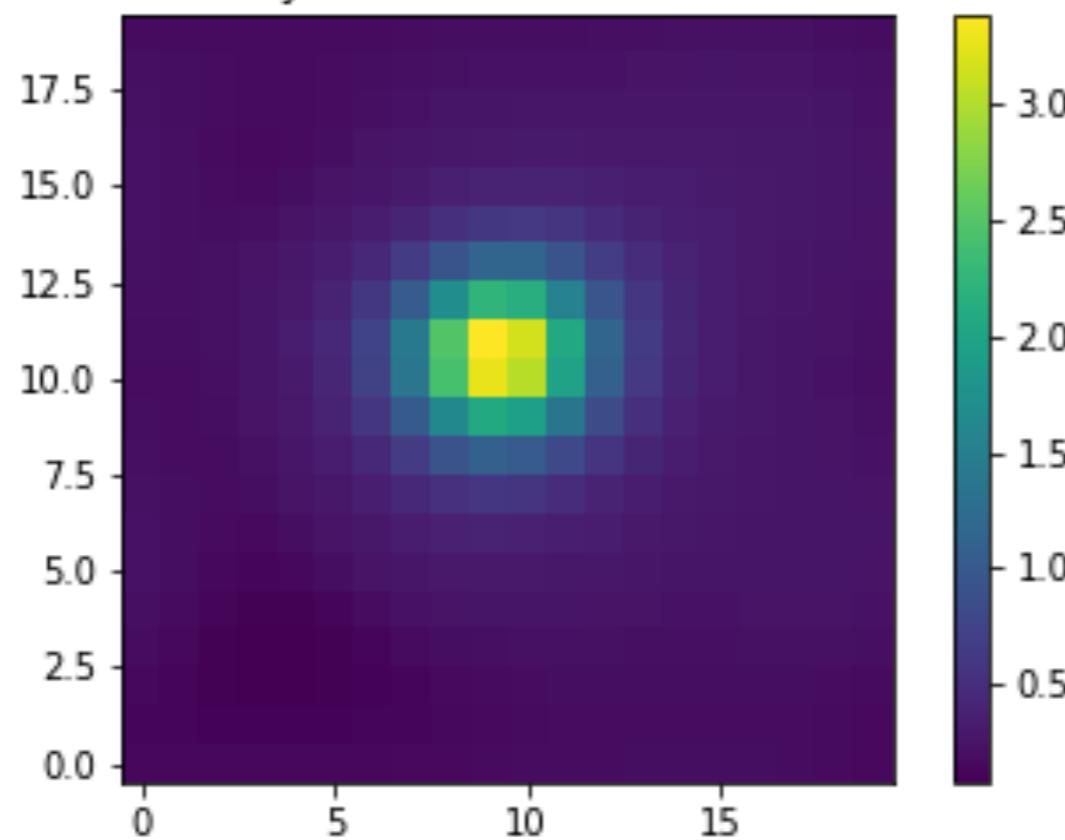
# Input



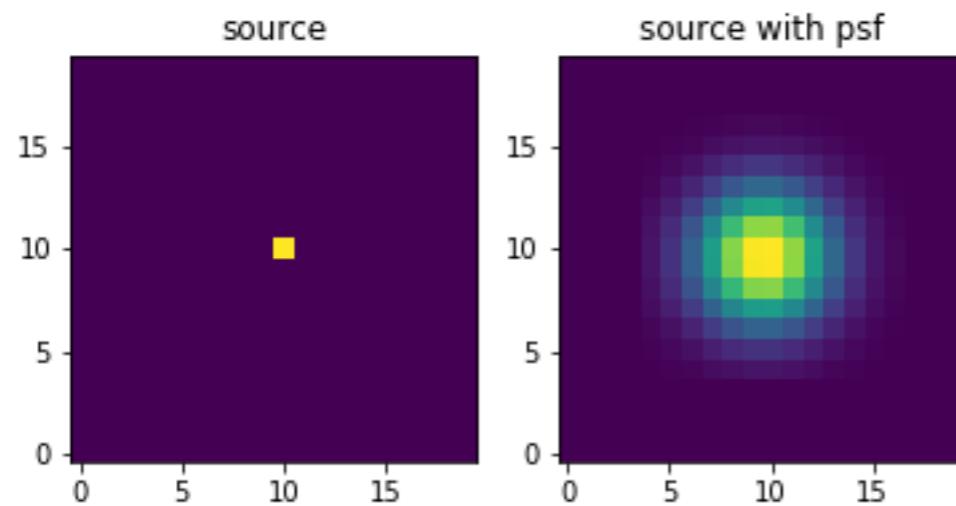
# Data



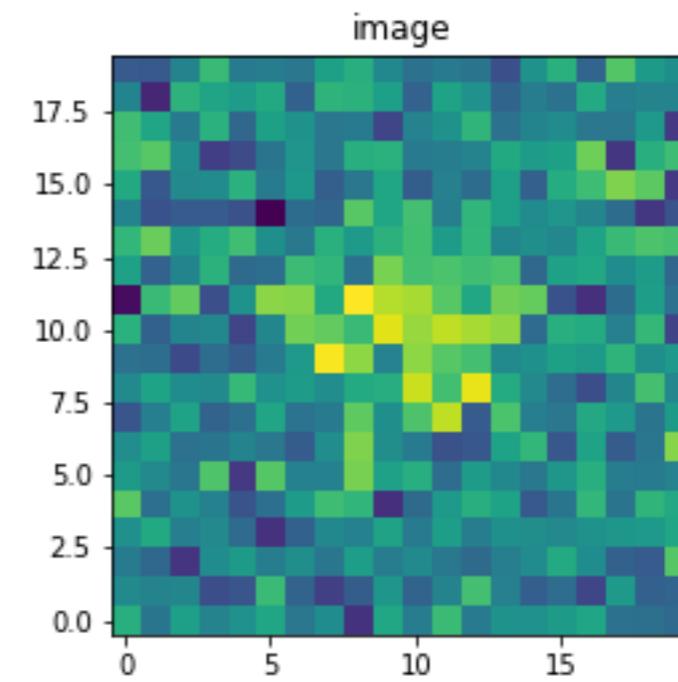
Bayes' factor for detection



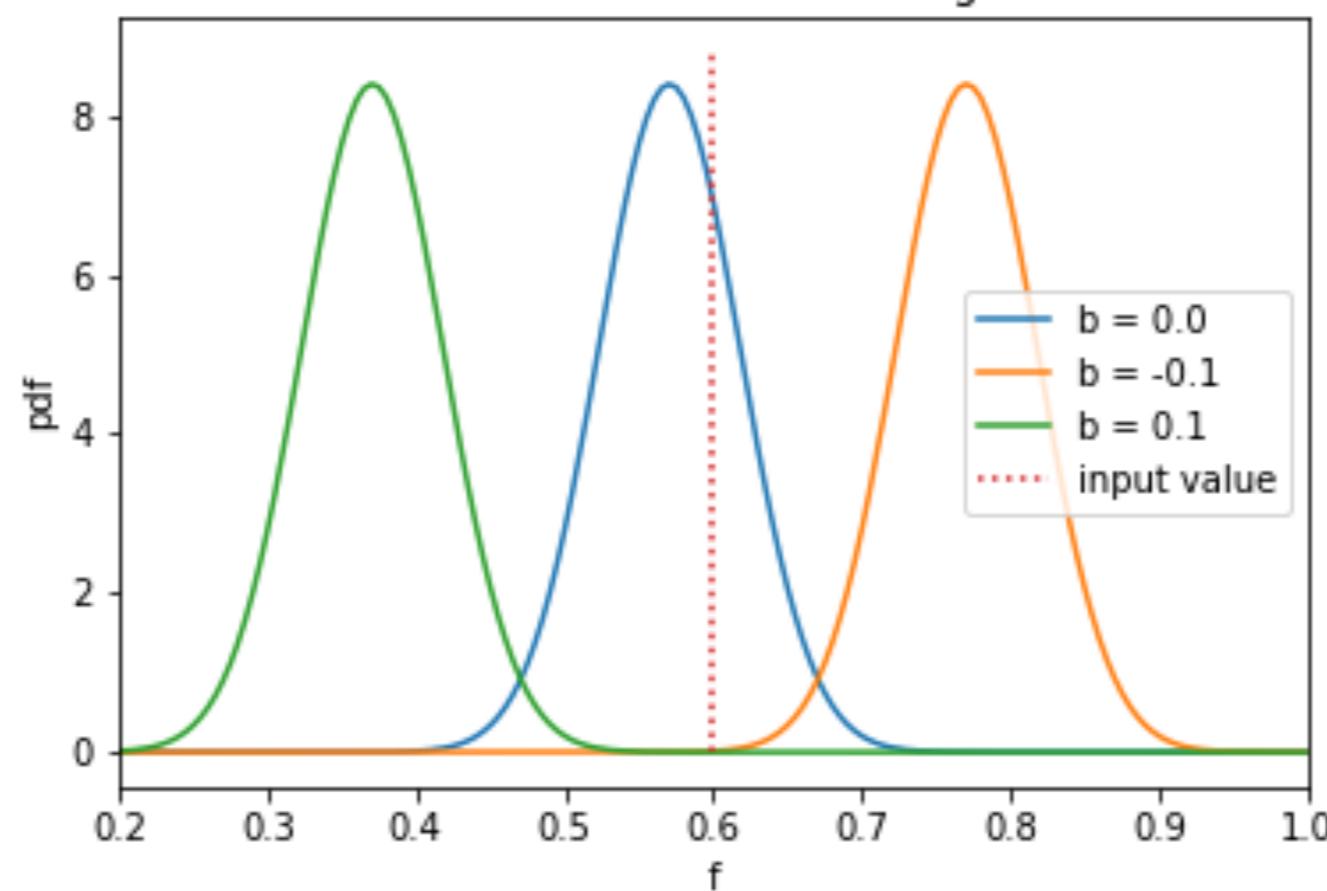
# Input



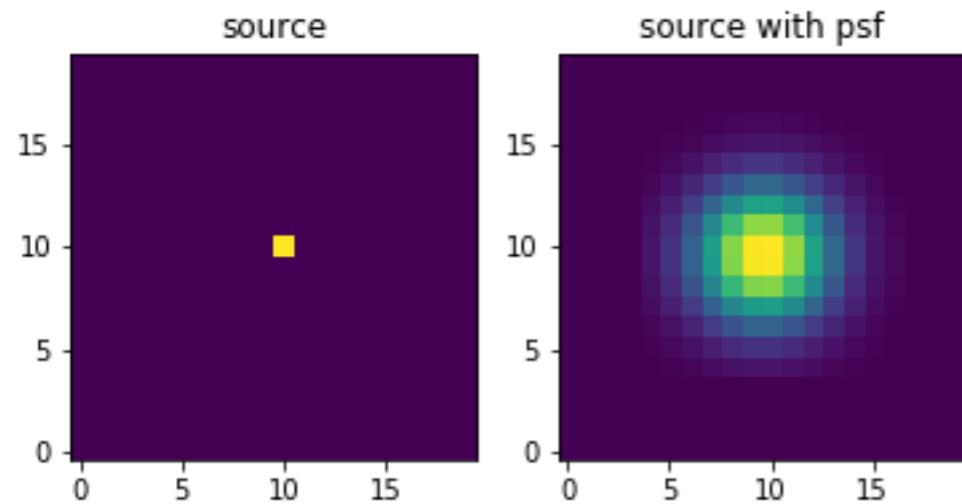
# Data



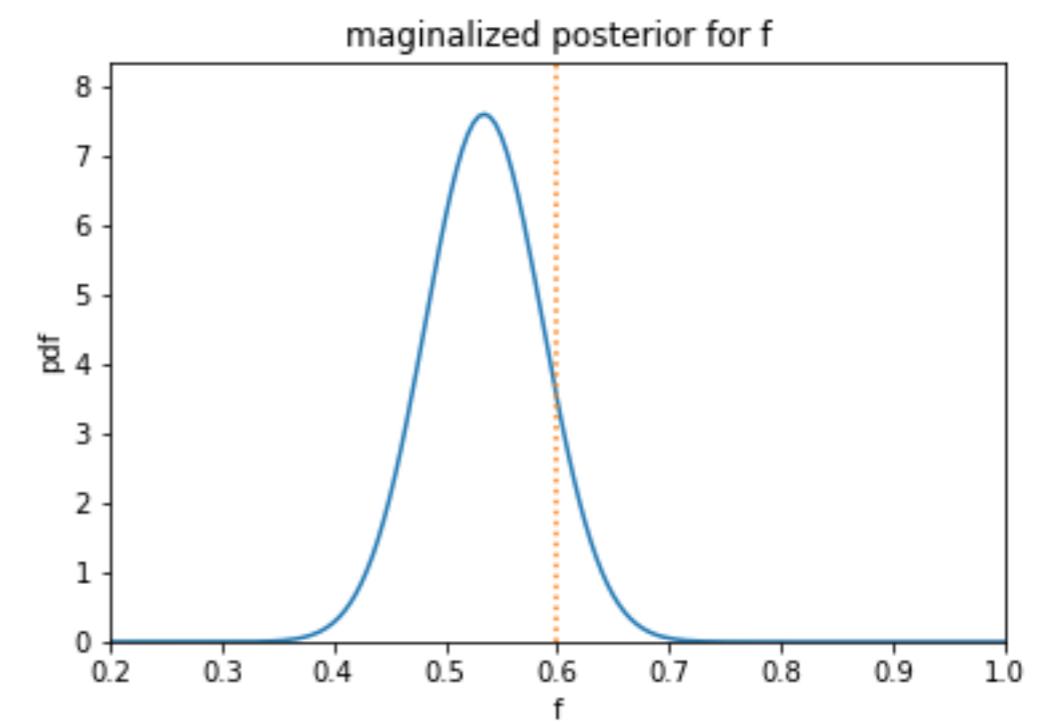
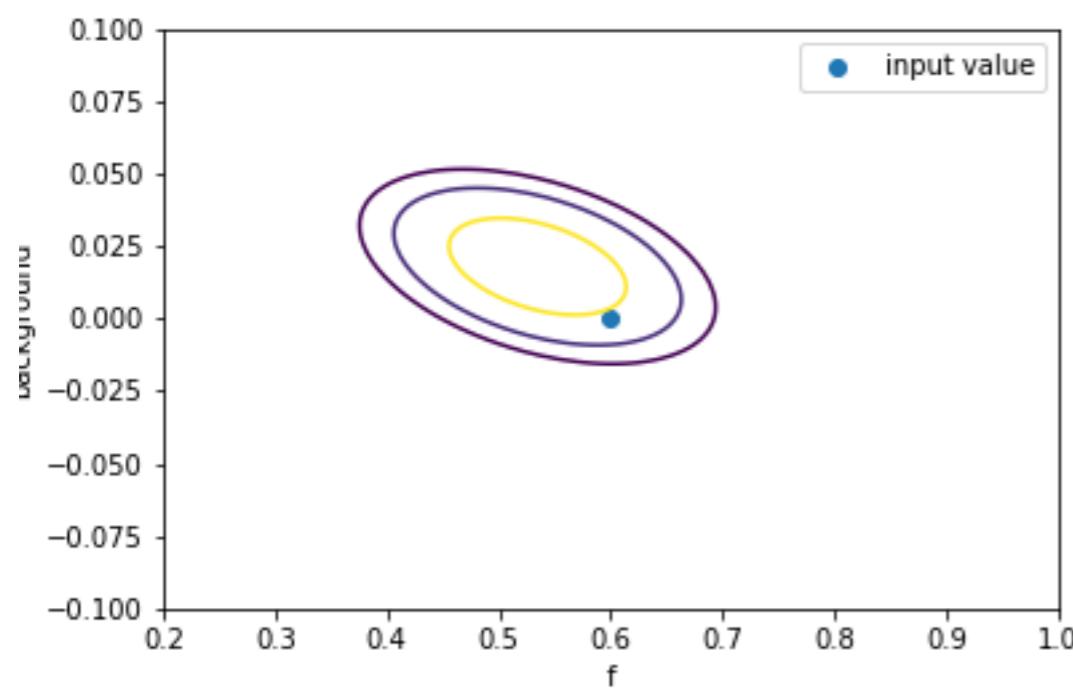
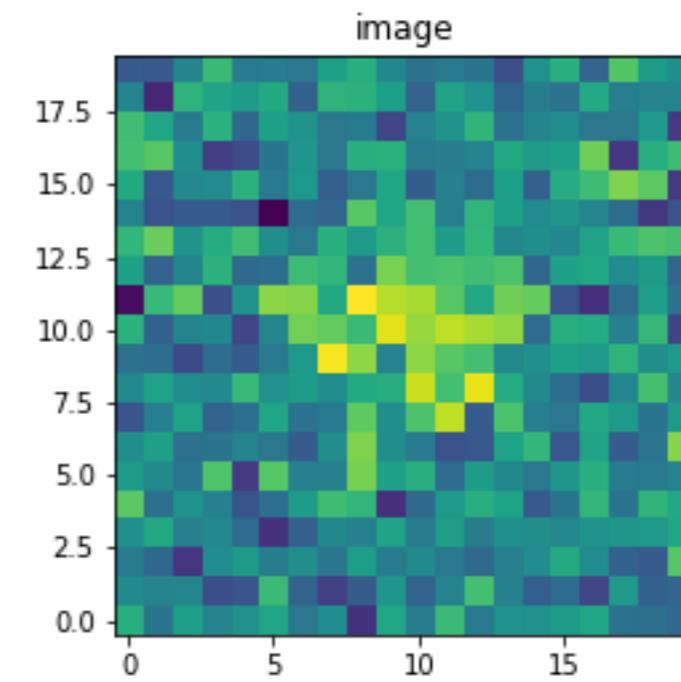
Posterior for  $f$  with fixed background



# Input



# Data



ignore the prior

$$B_{2,1} = \frac{\int d\theta \int d\beta \mathcal{L}(D|\theta, \beta) \pi(\theta, \beta)}{\int d\theta \mathcal{L}(D|\theta, \beta_o) \pi(\theta)} \underset{\pi(\hat{\theta}_{st})}{\simeq} \frac{\pi(\hat{\theta}_{ext}, \hat{\beta}_{ext})}{\pi(\hat{\theta}_{st})} \frac{\int d\theta d\beta \mathcal{L}(D|\theta, \beta)}{\int d\theta \mathcal{L}(D|\theta, \beta_o)}$$

and then the ratio of priors is simply dropped,

$$B_{2,1} \sim \frac{\int d\theta d\beta \mathcal{L}(D|\theta, \beta)}{\int d\theta \mathcal{L}(D|\theta, \beta_o)}.$$

# Bayesian Information Criterion

Consider the probability of the data given a model again

$$\begin{aligned} P(\mathbf{D}|M_i) &= \int_{-\infty}^{\infty} d\boldsymbol{\theta} \ \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}, M_i) \pi(\boldsymbol{\theta}|M_i) \\ &= \int_{-\infty}^{\infty} d\boldsymbol{\theta} \ e^{\ln \mathcal{L}(\mathbf{D}|\boldsymbol{\theta}, M_i)} \pi(\boldsymbol{\theta}|M_i) \end{aligned}$$

Expand likelihood around the minimum likelihood estimate (MLE)  $\hat{\boldsymbol{\theta}}$

$$\ln \mathcal{L}(\boldsymbol{\theta}, M_i) \simeq \ln \mathcal{L}(\hat{\boldsymbol{\theta}}, M_i) + \frac{1}{2} (\theta_i - \hat{\theta}_i) \left[ \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} \right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\theta_j - \hat{\theta}_j) + \dots$$

$$\mathcal{I}_{ij} \equiv - \left[ \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} \right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

If each of the  $n$  data points is independent then  $\mathcal{I}_{ij} = n \bar{\mathcal{I}}_{ij}$  where  $\bar{\mathcal{I}}_{ij}$  is the information per data point.

$$P(\mathbf{D}|M_i) \simeq \pi(\hat{\boldsymbol{\theta}}|M_i) \mathcal{L}(\hat{\boldsymbol{\theta}}, M_i) \frac{(2\pi)^{k/2}}{\sqrt{n^k |\bar{\mathcal{I}}|}}$$

where  $k$  is the number of parameters. Ignore  $|\bar{\mathcal{I}}|$

$$\text{BIC}_i \equiv k \ln n - 2 \ln [\mathcal{L}(\hat{\boldsymbol{\theta}}, M_i)]$$

so that

$$P(M_i|\mathbf{D}) \propto P(\mathbf{D}|M_i)P(M_i) \propto e^{-\text{BIC}/2} P(M_i).$$

If the BIC of two models differ by less than 2 there is considered to be no real reason to favor either one. If  $|\Delta\text{BIC}| = 2-6$  then there is some reason, 6 -10 is strong evidence and  $> 10$  is considered very strong evidence that one is better than the other.

# BP for linear model and Gaussian noise

Linear model and one independent variable  $\mathbf{f}(\mathbf{x}|\boldsymbol{\theta}) = \sum_i f^i(x)\boldsymbol{\theta}_i = \mathbf{f}_x \cdot \boldsymbol{\theta}$ .

So the mean (and mode) are what you might expect, the prediction using the best fit parameters.

$$\begin{aligned}\langle y^2 \rangle - \langle y \rangle^2 &= \int dy \int d\boldsymbol{\theta} y^2 \mathcal{L}(y|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X}) - \langle y \rangle^2 \\ &= \int d\boldsymbol{\theta} \left[ \sigma_y^2 + (\mathbf{f}(x) \cdot \boldsymbol{\theta})^2 \right] p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X}) - \langle y \rangle^2 \\ &= \int d\boldsymbol{\theta} \left[ \sigma_y^2 + \mathbf{f}(x) \boldsymbol{\theta} \boldsymbol{\theta}^T \mathbf{f}(x)^T \right] p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X}) - \langle y \rangle^2 \\ &= \sigma_y^2 + \mathbf{f}(x) \left[ (\mathbf{M}^T \mathbf{C}^{-1} \mathbf{M})^{-1} + \hat{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}^T \right] \mathbf{f}(x)^T - \langle y \rangle^2 \\ &= \sigma_y^2 + \mathbf{f}(x) (\mathbf{M}^T \mathbf{C}^{-1} \mathbf{M})^{-1} \mathbf{f}(x)^T\end{aligned}$$

# Numerical model selection / checking

The parametric bootstrap method discussed previously has some advantages, but its validity depends on asymptotic theorems which might not apply.

**posterior predictive p-values (PPP)** – The strategy is to use Bayesian prediction based on the observed data to generate mock data sets that can be used to calculate the cumulative distribution of a goodness-of-fit statistic.

$$p(\mathbf{x}) = \int d\boldsymbol{\theta} \ p(\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta}).$$

For the probability  $p(\boldsymbol{\theta})$  we can use the posterior given the observed data

$$p(\mathbf{x}) = \int d\boldsymbol{\theta} \ p(\boldsymbol{\theta}|\mathbf{d})p(\mathbf{x}|\boldsymbol{\theta}).$$

For any statistic  $T(x)$  we can use this to calculate its distribution

$$p(T) = \int d\theta \ p(\theta|\mathbf{d})p(T(x)|\theta).$$

and use it to do hypothesis testing.

In most cases this cannot be done analytically, but can be approximated with the following steps

- 1 Generate parameters set  $\theta_i$ ,  $i = 1 \dots N$  taken from the posterior  $p(\theta|\mathbf{d})$  using converged MCMC or another technique.
- 2 For each parameter set  $\theta_i$  generate a data set  $x_i$  taken from the likelihood  $p(x|\theta_i)$ . This is often a Gaussian, Poisson or another classical distribution that can be easily sampled from.
- 3 Calculate the statistic  $T_i = T(x_i)$  for all the  $x_i$ 's
- 4 Create the empirical cumulative distribution for the  $T_i$ 's to evaluate the p-value. In other words, the estimated p-value is

$$p = \frac{1}{N} \sum_i^N \Theta(T_i > T(\mathbf{d}))$$

# Some vocabulary

- $t(\mathbf{x})$  is a **sufficient statistic** for a parameter  $\theta$ . if it contains all the information in the data,  $\mathbf{d}$ , about that parameters. In this case the likelihood can be written

$$P(\mathbf{d}|\theta) = f(\mathbf{d})g(\theta, t(\mathbf{d})) \quad (438)$$

example: Gaussian with linear model parameters

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\theta}) \propto e^{-\frac{1}{2}X^2(\mathbf{x}, \boldsymbol{\theta})} \quad X^2(\mathbf{x}, \boldsymbol{\theta}) = X^2(\mathbf{x}, \hat{\boldsymbol{\theta}}(\mathbf{x})) + X^2(\hat{\boldsymbol{\theta}}(\mathbf{x}), \boldsymbol{\theta}) \quad (439)$$

- An estimator is **biased** if

$$\langle \hat{\boldsymbol{\theta}}(\mathbf{x}) \rangle \neq \boldsymbol{\theta}_o \quad (440)$$

where  $\boldsymbol{\theta}_o$  is the true value of the parameter.

- An estimator is **asymptotically unbiased** or **consistent** if

$$\lim_{n \rightarrow \infty} \langle \hat{\boldsymbol{\theta}}_n(\mathbf{x}) \rangle \rightarrow \boldsymbol{\theta}_o \quad (441)$$

where  $n$  is the number of data points.



# Numerical Methods

With a sufficiently large sample drawn from a distribution one can use the **law of large numbers** to estimate any expectation value

$$E[g(\mathbf{x})] = \int_{-\infty}^{\infty} d^n \mathbf{x} p(\mathbf{x}) g(\mathbf{x}) \simeq \frac{1}{n} \sum_i g(\mathbf{x}_i)$$

where the  $\mathbf{x}_i$ 's are drawn from the distribution  $p(\mathbf{x})$ .

# Numerical confidence / credibility Levels

**simulation / theory -> data -> statistic**

How many samples are needed?

If the probability of a statistic being larger than  $X$  is  $p$  then we know that the probability of  $k$  samples out of  $n$  being larger than  $X$  is given by the binomial distribution

$$P(k|p, n) = \binom{n}{k} p^k (1-p)^{n-k}$$

Assuming a uniform prior and using the integral form of the beta function we can renormalize this to get the posterior for  $p$

$$\begin{aligned} P(p|n, k) &= \frac{\Gamma(n+2)}{\Gamma(n-k+1)\Gamma(k+1)} p^k (1-p)^{n-k} \\ &= \frac{(n+1)!}{(n-k)!k!} p^k (1-p)^{n-k} \end{aligned}$$

- Simulate a data sets with the null hypothesis and calculate the statistic of interest,  $n$  times.
- The posterior for the p-value at the observed value is:

$$P(p|n, k) = \frac{(n+1)!}{(n-k)!k!} p^k (1-p)^{n-k}$$

where  $k$  is the number of simulations with a smaller value for the statistic.

- The variance of this distribution is

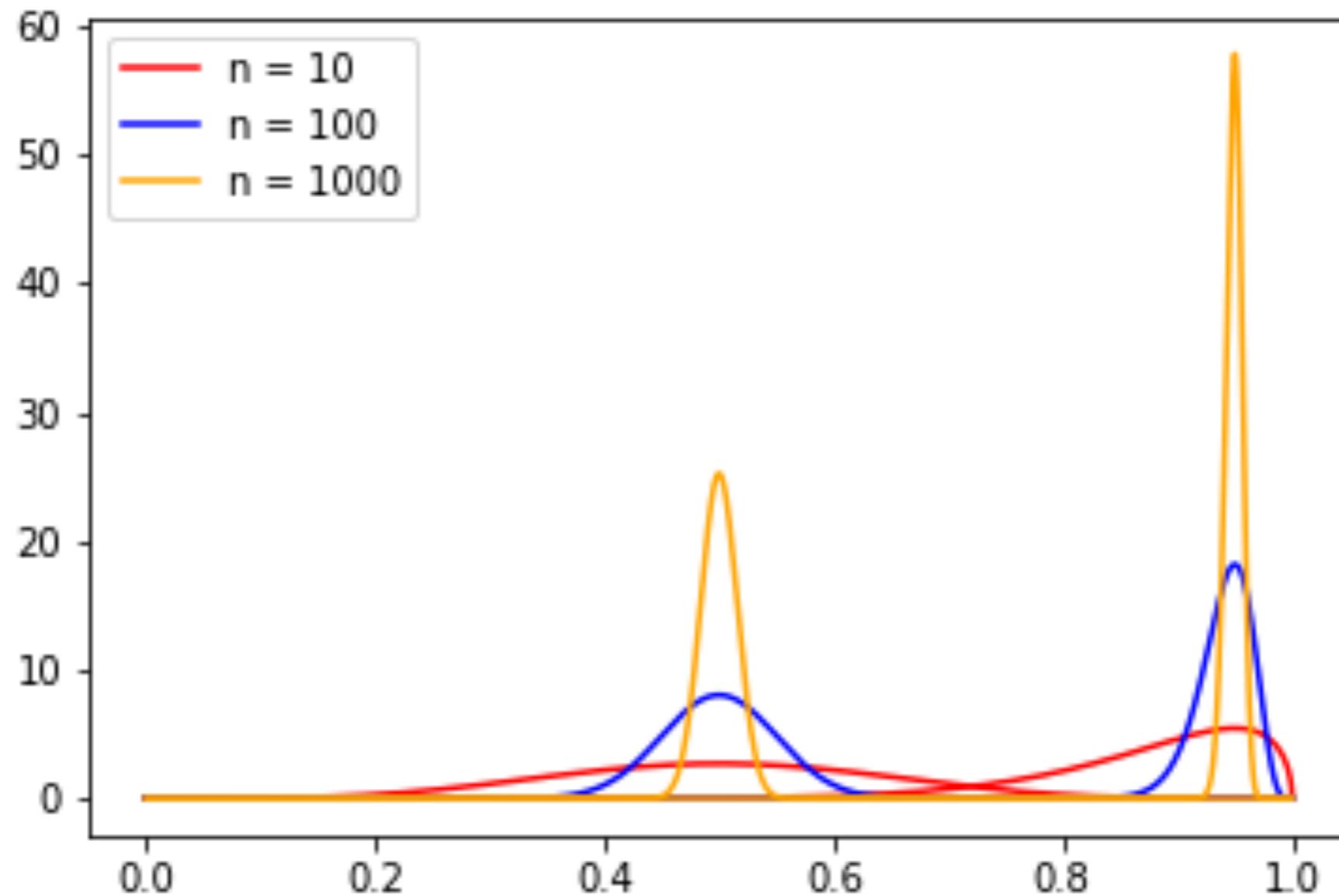
$$\begin{aligned}\sigma_p^2 &= \frac{(k+1)}{(n+2)} \left[ \frac{(k+2)}{(n+3)} - \frac{(k+1)}{(n+2)} \right] \\ &\approx \frac{3\langle p \rangle (1 - \langle p \rangle)}{n} + \mathcal{O}(1/n^2)\end{aligned}$$

If you want to know the p-value to an accuracy of 0.001 you will need

$$n \gtrsim \frac{3\langle p \rangle (1 - \langle p \rangle)}{(0.001)^2} = 3 \times 10^6 \langle p \rangle (1 - \langle p \rangle)$$

simulations. For  $\langle p \rangle = 0.99$  this is  $n \gtrsim 3 \times 10^4$ .

## Posteriors for the quantile of a point given a Monte Carlo sample from the distribution



# Monte Carlo Integration

$$\begin{aligned}\int_{\partial V} d^n x \, g(\mathbf{x}) &= \int_{\partial V} d^n x \, p(\mathbf{x}) \left( \frac{g(\mathbf{x})}{p(\mathbf{x})} \right) \\ &= \int_{-\infty}^{\infty} d^n x \, p(\mathbf{x}) \left( \frac{g(\mathbf{x})}{p(\mathbf{x})} \right) \Theta(\mathbf{x} \in V)) \\ &= \left\langle \left( \frac{g(\mathbf{x})}{p(\mathbf{x})} \right) \Theta(\mathbf{x} \in V) \right\rangle \\ &\simeq \frac{1}{n} \sum_{x_i \in V} \left( \frac{g(\mathbf{x}_i)}{p(\mathbf{x}_i)} \right) \quad \mathbf{x}_i \sim p(\mathbf{x})\end{aligned}$$

The estimated error on this would be

$$\pm \frac{1}{n} \sqrt{\sum_{x_i \in V} \left( \frac{g(\mathbf{x}_i)}{p(\mathbf{x}_i)} \right)^2 - \frac{1}{n} \left( \sum_{x_i \in V} \frac{g(\mathbf{x}_i)}{p(\mathbf{x}_i)} \right)^2}$$

## Importance sampling in Bayesian inference

Sampling from the distribution  $q(\theta)$  an estimate of the expectation of any function  $f(\theta)$  over the posterior is

$$\begin{aligned} E[f(\theta)] &= \int_{-\infty}^{\infty} d\theta \ p(\theta|\mathbf{d})f(\theta) \\ &\simeq \frac{\sum_i^n w_i f(\theta_i)}{\sum_i^n w_i} \quad \text{where } w_i = \frac{\mathcal{L}(\mathbf{d}|\theta_i)\pi(\theta_i)}{q(\theta_i)} \\ &\equiv \hat{\mu}_f \end{aligned}$$

An estimate of the variance of this is

$$\sigma_{\hat{\mu}_f}^2 = \frac{\sum_i^n w_i^2 (f(x_i) - \hat{\mu}_f)^2}{[\sum_i^n w_i]^2}$$

which can be monitored until the desired accuracy is obtained.

# Markov Chains

A **chain** is an ordered series of random numbers,  $x_1 \dots x_n \dots$  where the conditional probability of each element given the other elements is specified –  $p(x_n|x_1 \dots)$ .

- A **Markov chain** is a chain where the conditional probability of any element  $x_n$  can be expressed as a function of only the previous element  $x_{n-1}$ . (The future depends only on the present and not on the past, although the present does depend on the past.)
- The probability  $p(x_{n+1}|x_n)$  is known as the Markov chain's **transition kernel**.

To be **ergodic** the chain must be:

- 1 *irreducible* - A chain starting at any state  $x_0$  can reach any other state after a finite number of steps, not necessarily 1 step.
- 2 *aperiodic* - The chain will not return to the same state after some fixed number of steps and all multiples of this number of steps.
- 3 *positive recurrent* - The expectation value for the number of steps between any two states is finite.

It is also true that a Markov chain is ergodic if there is a number  $N$  such that any state can be reached from any other state in  $N$  steps and any number of steps larger than  $N$ .

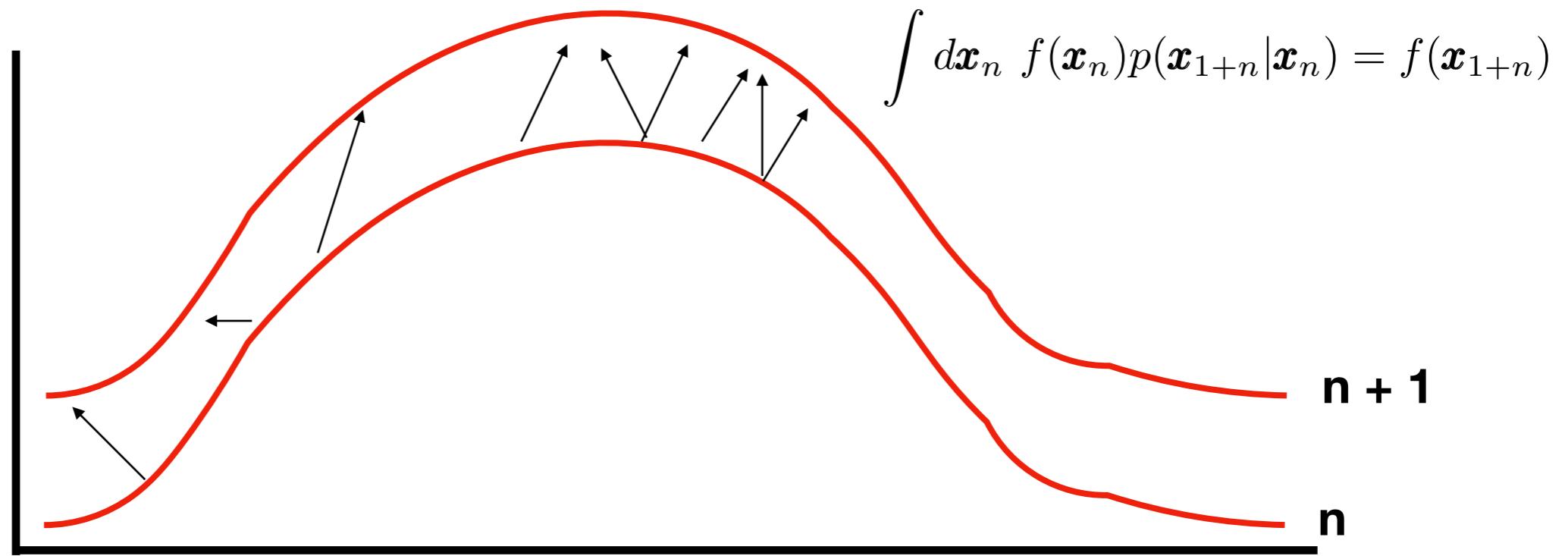
The most important consequence of ergodicity is that the chain has a unique **stationary distribution**  $f(\mathbf{x})$  such that

$$\int_{-\infty}^{\infty} d\mathbf{x}_n f(\mathbf{x}_n) p(\mathbf{x}_{1+n} | \mathbf{x}_n) = f(\mathbf{x}_{1+n}) \quad (18)$$

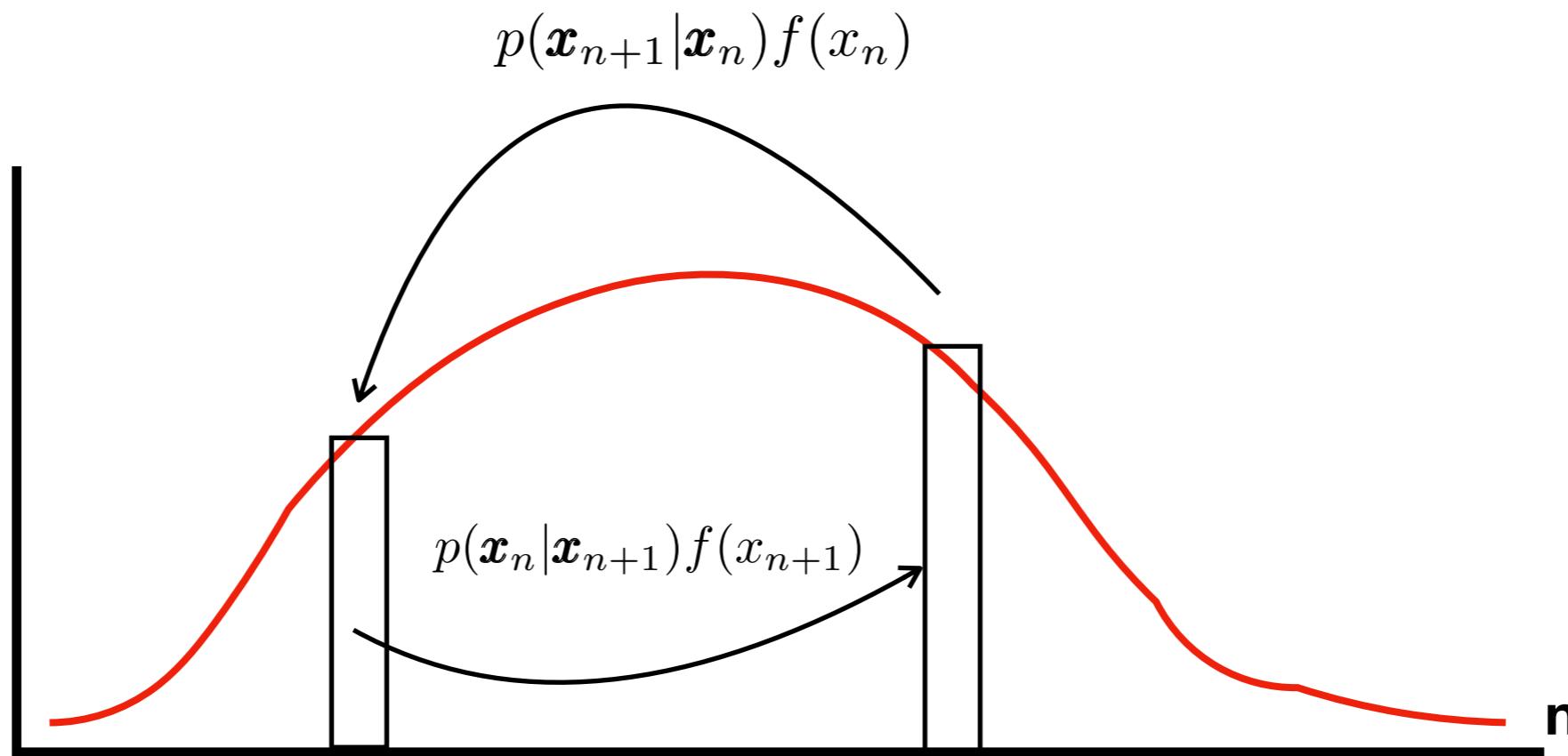
which means that we can produce chains whose states are distributed according to  $f(\mathbf{x})$  if we can find a transition kernel that satisfies this requirement. And the law of large numbers will apply

$$E[g(\mathbf{x})] = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^N g(\mathbf{x}_n) \quad (19)$$

stationary  
distribution



detailed  
balance



The kernel satisfies **detailed balance** if:

$$p(\mathbf{x}_{n+1} | \mathbf{x}_n) f(\mathbf{x}_n) = p(\mathbf{x}_n | \mathbf{x}_{n+1}) f(\mathbf{x}_{n+1})$$

A transition kernel that satisfies detailed balance will have  $f(\mathbf{x})$  as a stationary distribution.

## The Metropolis-Hastings algorithm :

Starting at state  $\mathbf{x}_n$

- 1 Choose a new trial point  $\mathbf{x}_t$  from a **proposal distribution**  $q(\mathbf{x}_t|\mathbf{x}_n)$ .
- 2 Calculate

$$\alpha(\mathbf{x}_t, \mathbf{x}_n) = \min \left\{ 1, \frac{q(\mathbf{x}_n|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_n)} \frac{f(\mathbf{x}_t)}{f(\mathbf{x}_n)} \right\}$$

- 3 If  $\alpha < 1$  draw a uniform deviate between 0 and 1. If  $\alpha$  is large than this number accept the trial state and set  $\mathbf{x}_{n+1} = \mathbf{x}_t$ . Otherwise set  $\mathbf{x}_{n+1} = \mathbf{x}_n$ . In other words, accept the trial state with probability  $\alpha(\mathbf{x}_t, \mathbf{x}_n)$ .
- 4 repeat

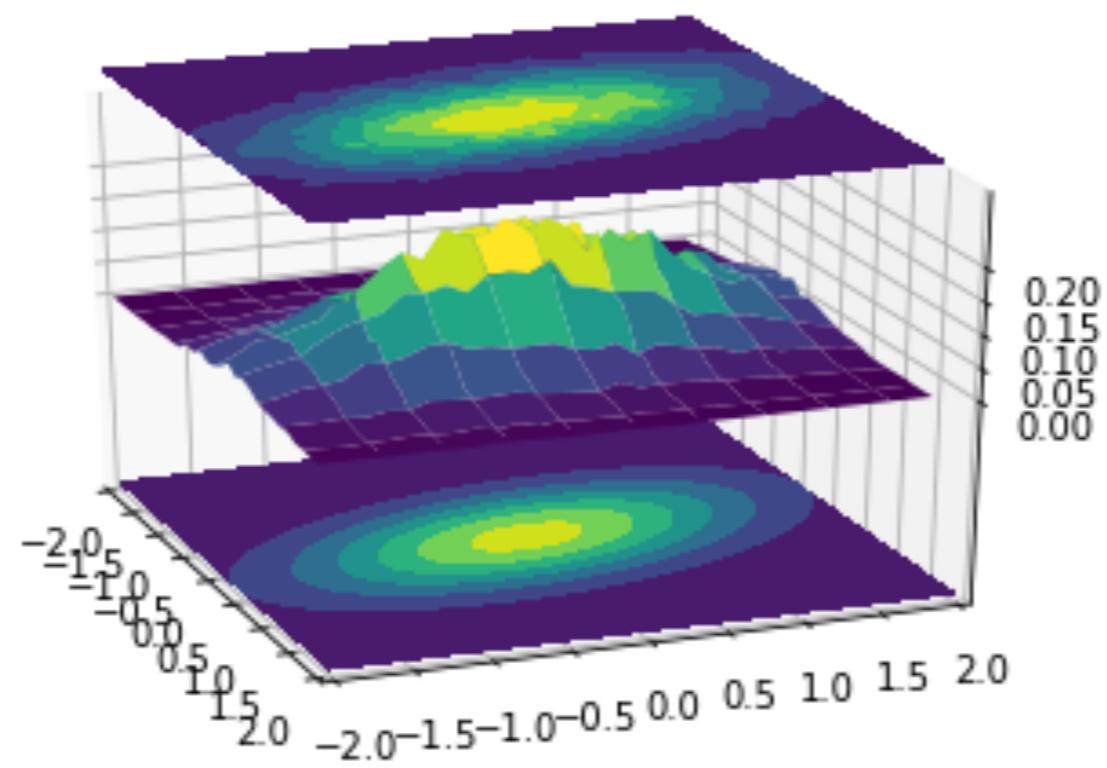
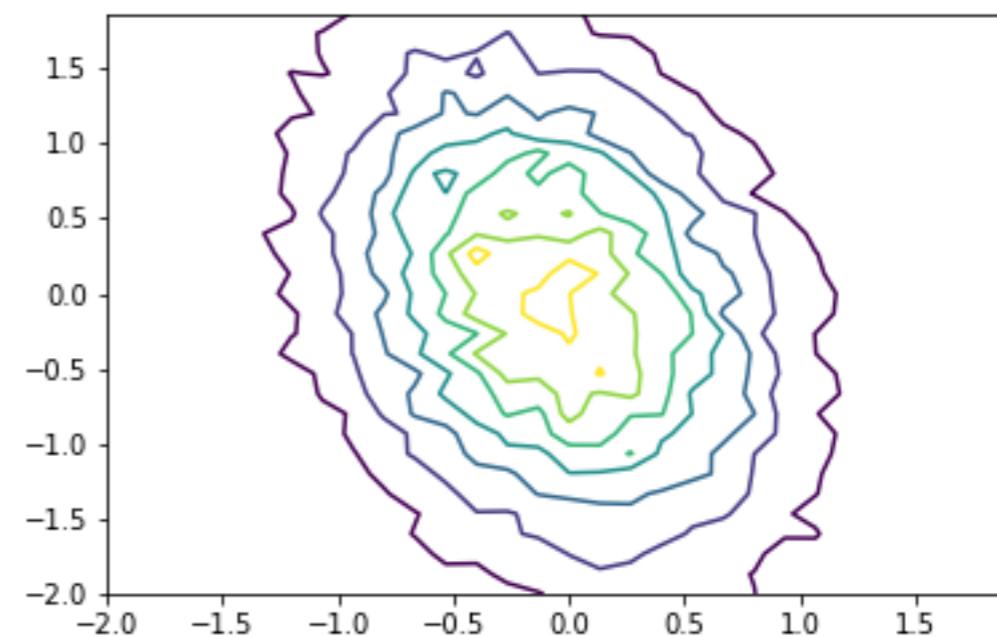
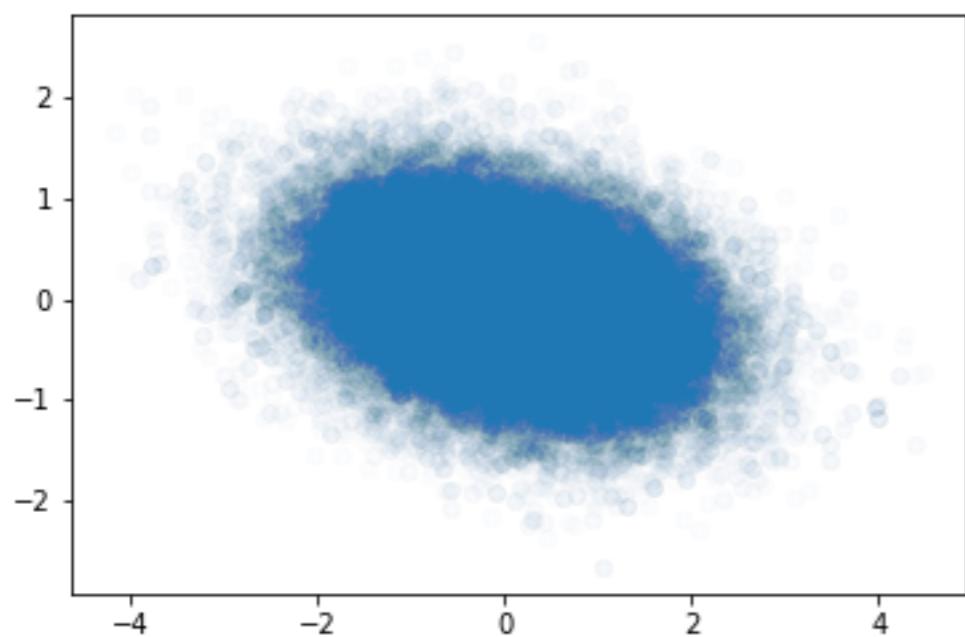
In this case we can easily see how detail balance is satisfied by this algorithm

$$\begin{aligned} p(\mathbf{x}_{n+1}|\mathbf{x}_n)f(\mathbf{x}_n) &= q(\mathbf{x}_{n+1}|\mathbf{x}_n)\alpha(\mathbf{x}_{n+1}, \mathbf{x}_n)f(\mathbf{x}_n) \\ &= \begin{cases} q(\mathbf{x}_n|\mathbf{x}_{n+1})f(\mathbf{x}_{n+1}) & , \quad q(\mathbf{x}_n|\mathbf{x}_{n+1})f(\mathbf{x}_{n+1}) < q(\mathbf{x}_{n+1}|\mathbf{x}_n)f(\mathbf{x}_n) \\ q(\mathbf{x}_{n+1}|\mathbf{x}_n)f(\mathbf{x}_n) & , \quad q(\mathbf{x}_n|\mathbf{x}_{n+1})f(\mathbf{x}_{n+1}) > q(\mathbf{x}_{n+1}|\mathbf{x}_n)f(\mathbf{x}_n) \end{cases} \end{aligned}$$

and

$$p(\mathbf{x}_n|\mathbf{x}_{n+1})f(\mathbf{x}_{n+1}) = q(\mathbf{x}_n|\mathbf{x}_{n+1})\alpha(\mathbf{x}_n, \mathbf{x}_{n+1})f(\mathbf{x}_{n+1})$$

which will be the same as above in the same cases so detailed balance is satisfied.



Although the MCMC is guaranteed to converge under the conditions mentioned above, it might take a *very long time*.

To achieve good *mixing* the **rejection rate** of proposed moves must not be too high or too low. If it is too high the chain will have many duplicated points that will not fill parameter space in an even way. If the rejection rate is too low the chain will move, but not fast enough to get around the space. A rule of thumb is that you want a rejection rate of about 80%, i.e. an acceptance rate of 20%. This rate can be changed by adjusting the proposal function  $q(\mathbf{x}_t | \mathbf{x}_n)$ .

A popular choice is the multivariate Gaussian centered on the current point so  $\mathbf{x}_{n+1} = \mathbf{x}_n + \mathbf{y}$  where  $\mathbf{y} \sim \mathcal{N}(0, \sigma)$  is samples from a multivariate Gaussian. But the  $\sigma$ 's (or the covariance matrix  $\mathbf{C}$ ) is not specified. These variances need to be adjusted until an acceptable rejection rate is found.

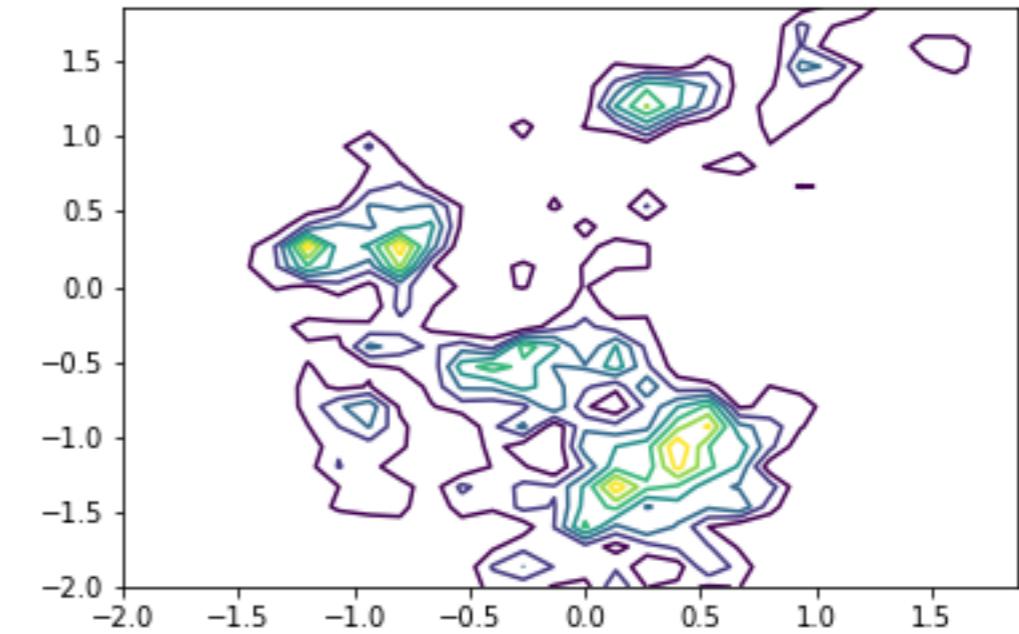
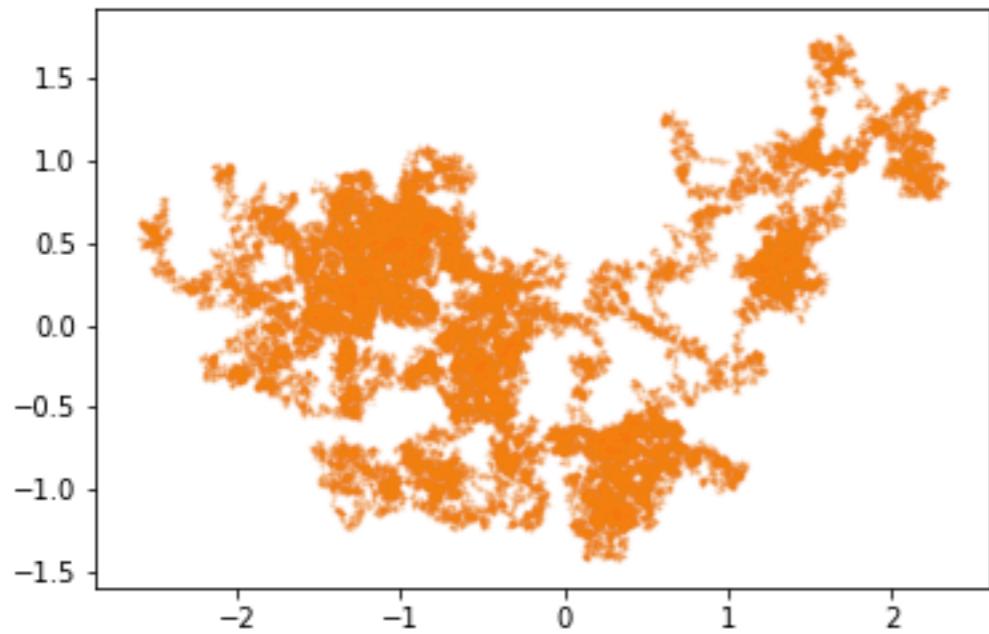
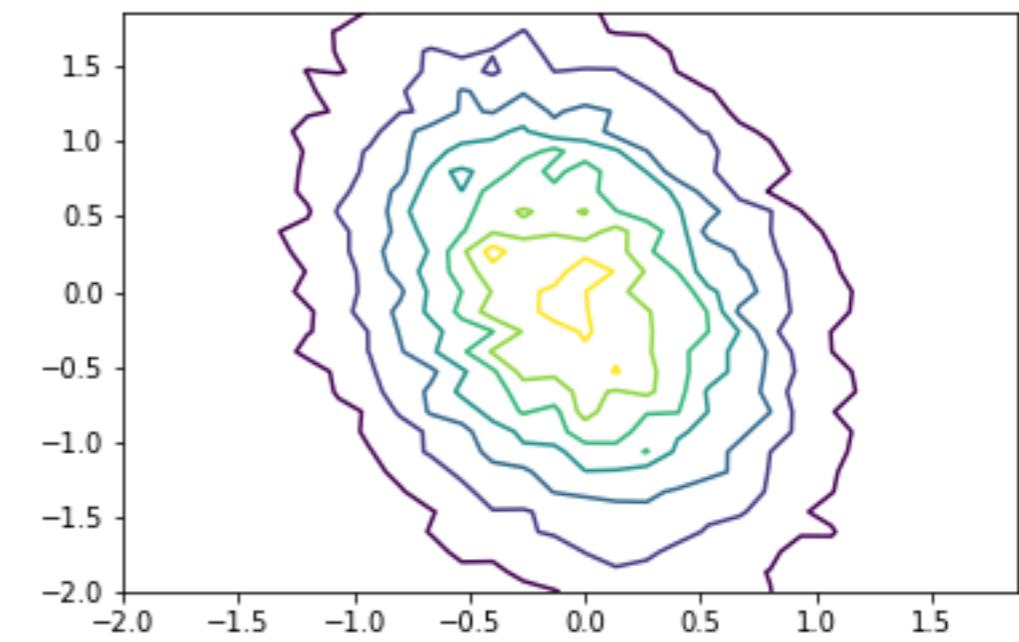
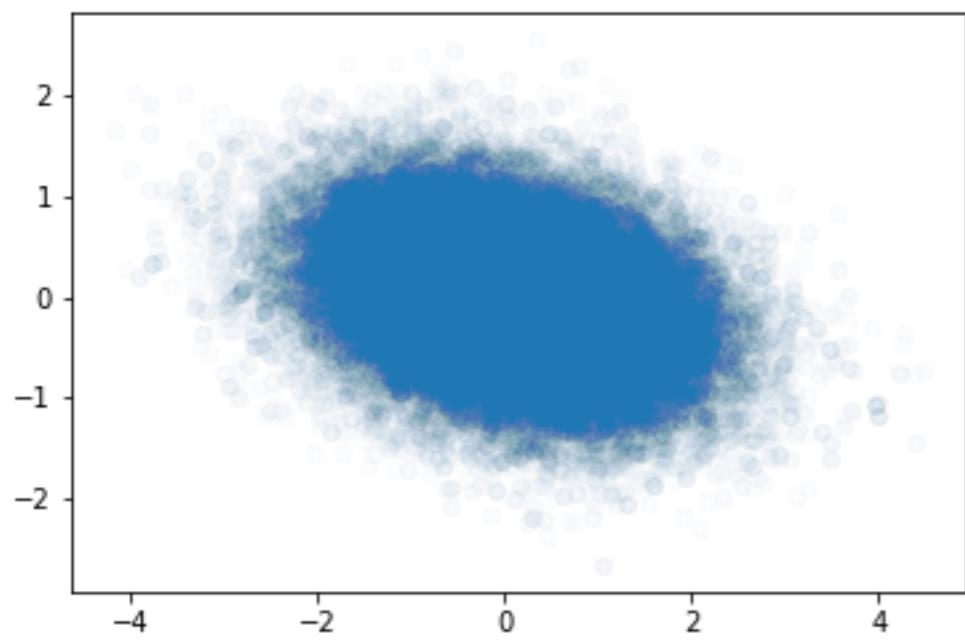
To initialize the chain one must guess a point in parameter space. This will usually not be a place of high probability without prior knowledge.

During the **burn in period** the chain is not near its stationary distribution. This part of the chain is usually discarded. There is no perfect method for determining how long the burn in period should be.

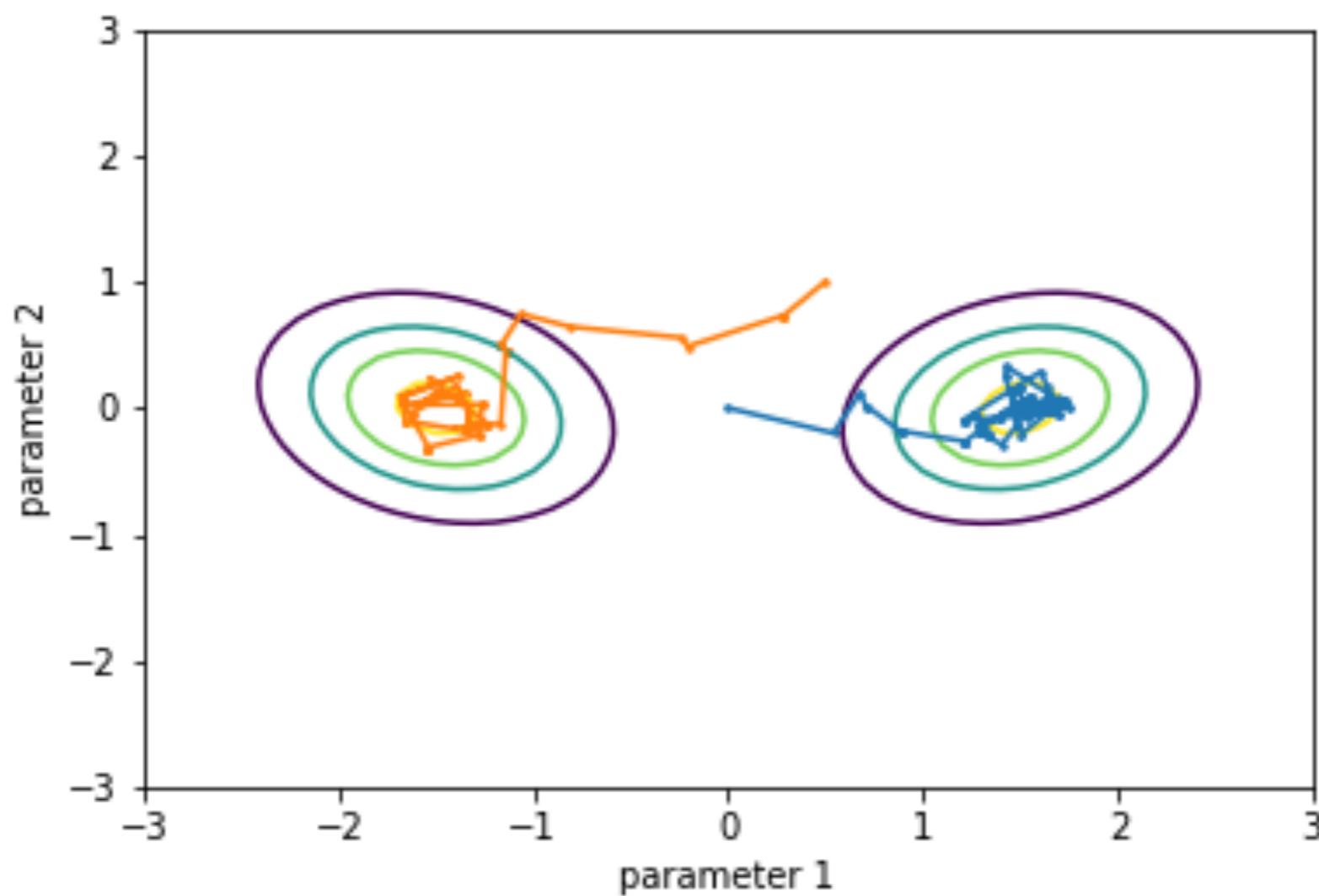
Some times the maximum of the distribution can be found by some minimization technique and then MCMC is used to map the posterior to find variances and covariances.

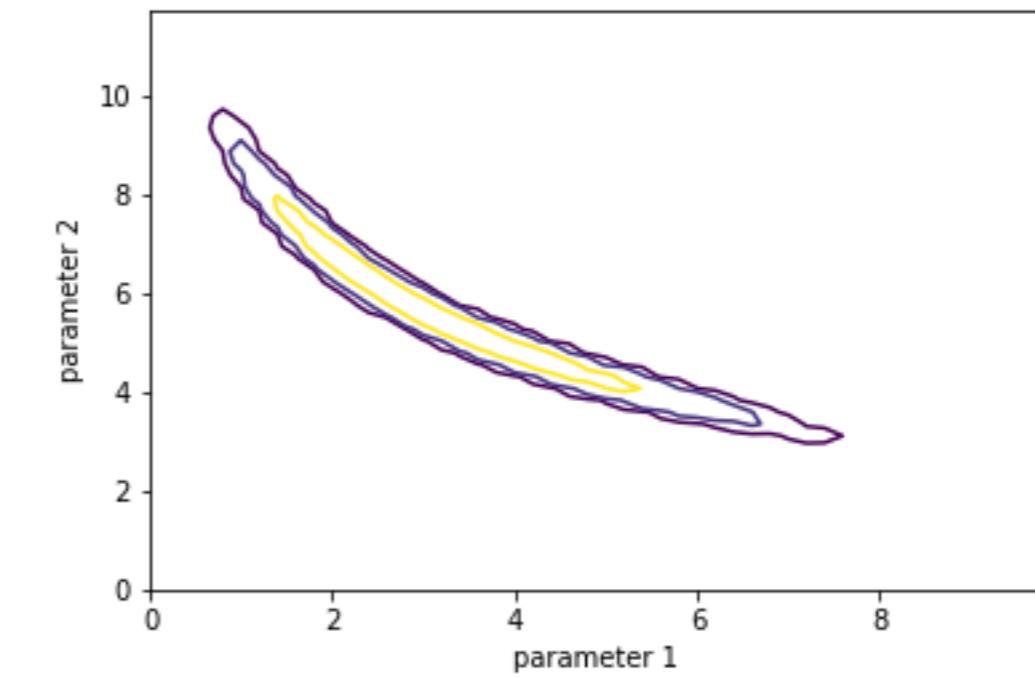
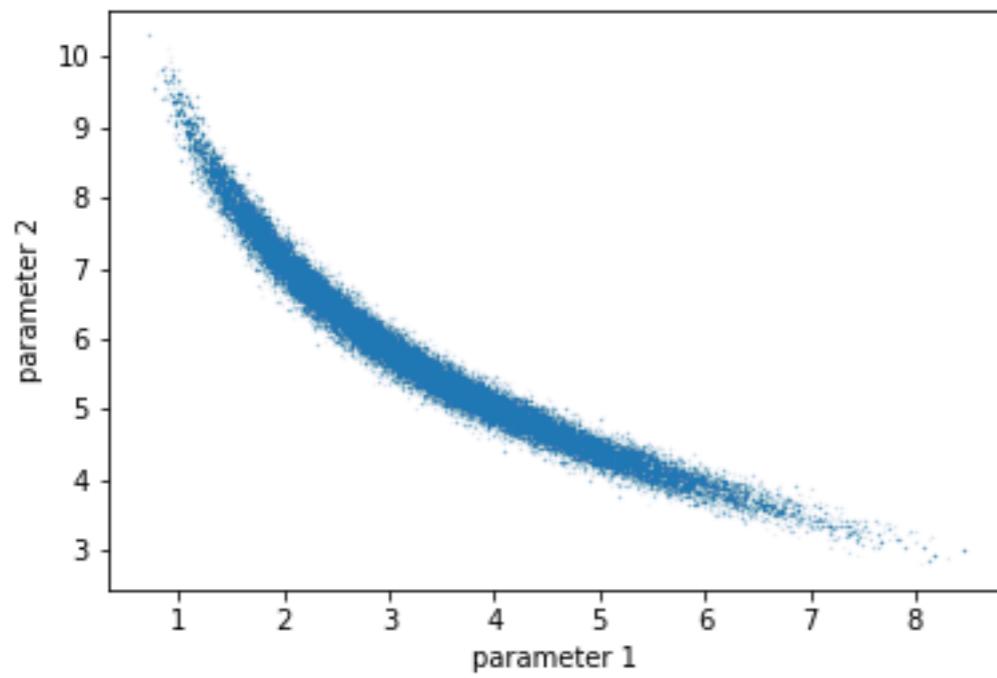
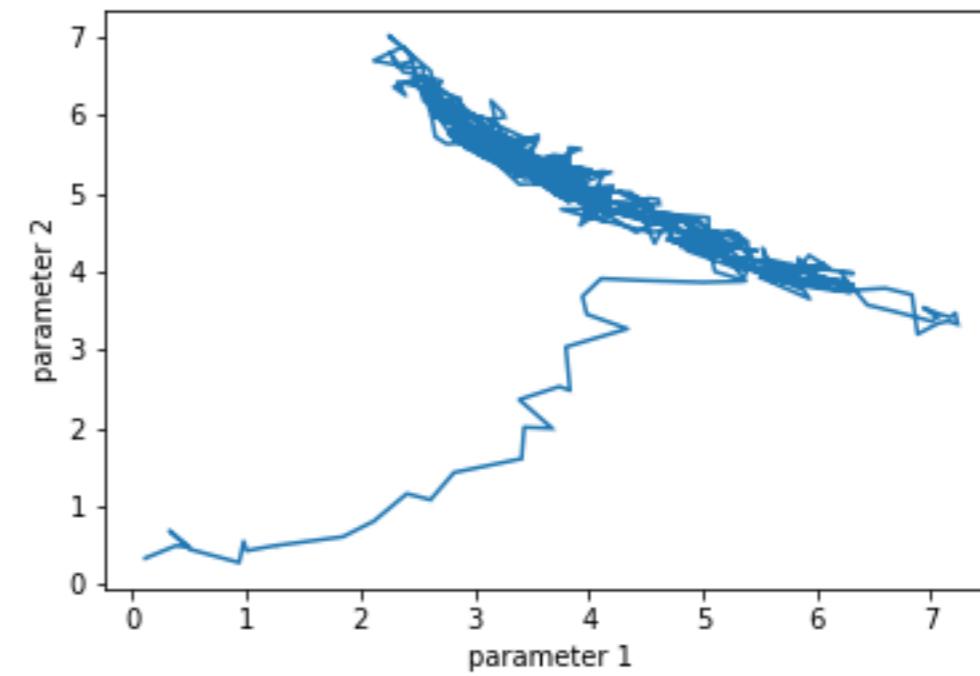
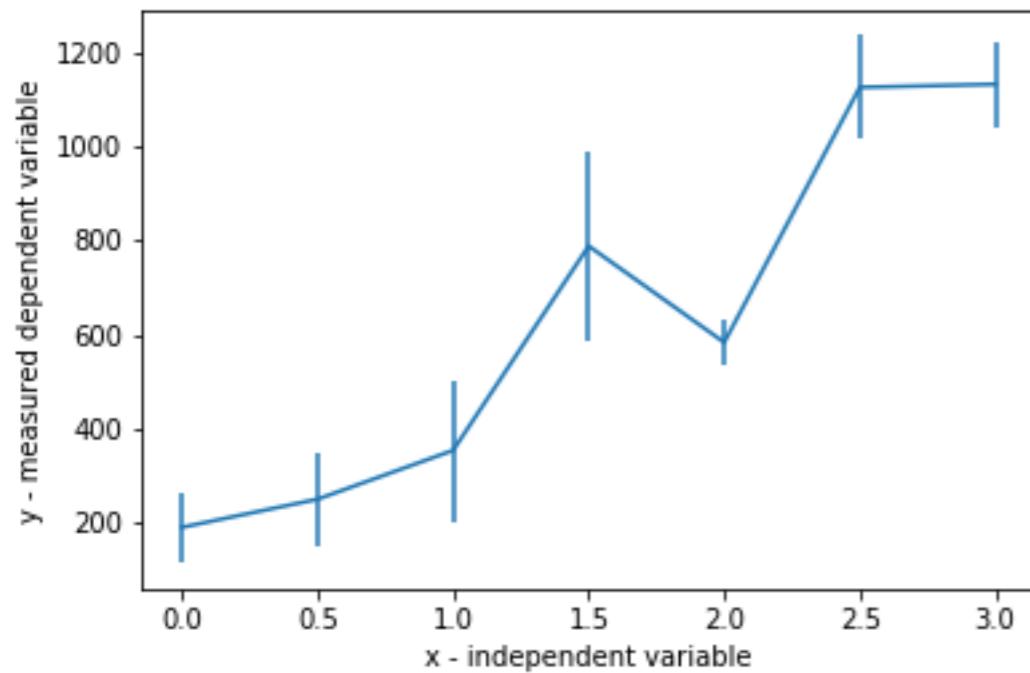
The most difficult cases for MCMC arise when:

- The **initial guess** is so far from any peaks and the probability is so flat out there that the chain never finds a peak. It is sometimes the case that in low probability regions the calculation of  $f(\mathbf{x})$  has a numerical underflow error or is dominated by numerical noise in which case the chain may wander around without getting anywhere.
- The **parameters are degenerate**. Imagine a  $f(\mathbf{x})$  that has a narrow ridge. If the proposal distribution is isotropic it will be either too wide in one direction so that the rejection rate is too high or it will be too small and creep along the ridge very slowly. In the case of a linear degeneracy you might be able learn something about the distribution and then make your proposal distribution anisotropic in a way that improves convergence. A nonlinear degeneracy is much more of a problem.
- The distribution has **multiple modes**. This is probably the hardest problem to deal with. If there are multiple peaks in the distribution that are separated by regions of low probability then the chain can easily get caught in one peak where its probability of transitioning to the other is very small.



**proposal function is too small**





# convergence of the chain

Unfortunately there is no fool proof ways of knowing when the chain has converged.

The autocorrelation for each of the parameters as a function of the *lag*,  $m$ , the separation in the chain. It can be defined as

$$C_{\alpha,\beta}(m) = \frac{\sum_{i=1}^{N-m} (\alpha_i - \bar{\alpha})(\beta_{i+m} - \bar{\beta})}{\sqrt{\left(\sum_{i=1}^{N-m} (\alpha_i - \bar{\alpha})^2\right) \left(\sum_{i=m}^N (\beta_i - \bar{\beta})^2\right)}}$$

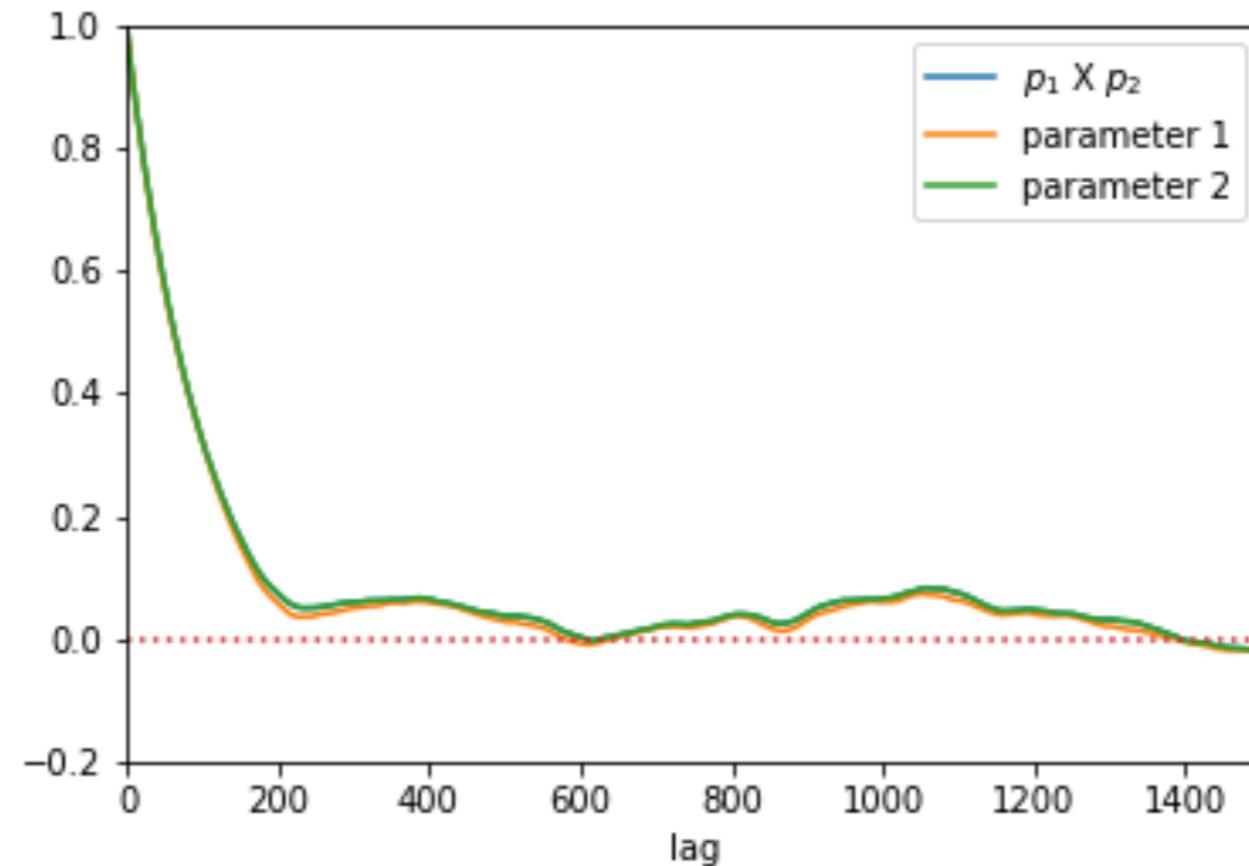
Distant points along the chain should not be correlated so this function should oscillate about zero for large lag,  $m$ . The first time this function drops to zero or near zero is an estimate of the **correlation length**,  $N_{corr}$ .

Effective number of independent samples in the chain as

$$N_{eff} = \frac{N_{chain}}{N_{corr}}$$

The difference between the 95% and 99% contour levels depend on only 4% of the particles. This should not be smaller than  $N_{corr}$ .

# convergence of the chain



**Figure:** The correlation coefficient as a function of lag in the MCMC chain. Shown are the autocorrelation for the two parameters and the crosscorrelation between them.

## Gelman-Rubin diagnostic $\hat{R}$ for convergence of a chain

If we have  $m$  independent chains each of length  $n$  and  $\theta_i^\alpha$  is the  $i$ th parameter value of the  $\alpha$ th chain we can define the following quantities:

$$\bar{\theta}^\alpha = \frac{1}{n} \sum_i^n \theta_i^\alpha$$

$$\bar{\bar{\theta}} = \frac{1}{m} \sum_\alpha^m \bar{\theta}^\alpha$$

$$s_\alpha^2 = \frac{1}{n-1} \sum_i^n (\theta_i^\alpha - \bar{\theta}^\alpha)^2$$

$$B = \frac{n}{m-1} \sum_\alpha^m (\bar{\theta}^\alpha - \bar{\bar{\theta}})^2$$

$$W = \frac{1}{m} \sum_\alpha^m s_\alpha^2$$

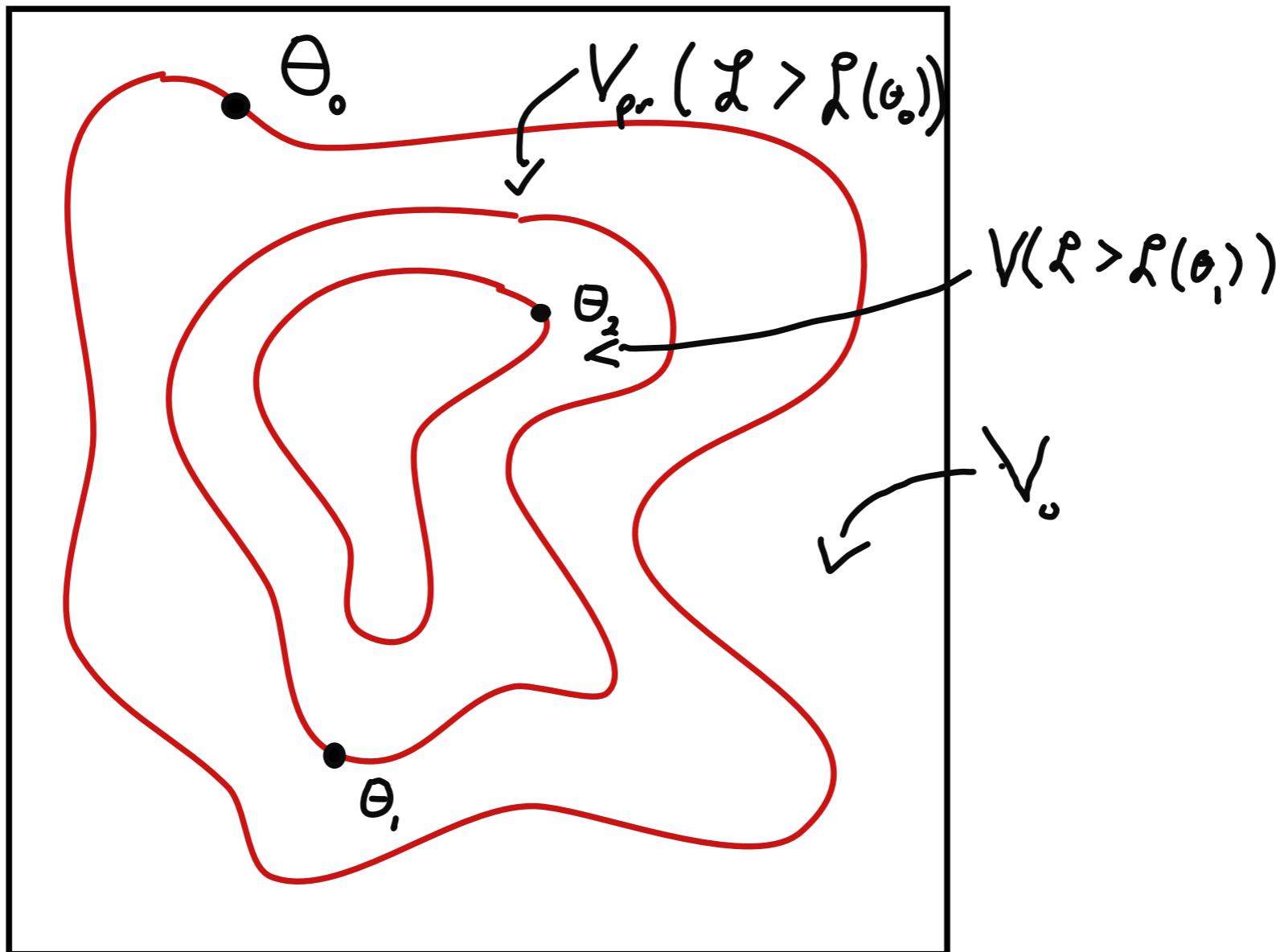
$$V = \frac{n-1}{n} W + \frac{M+1}{nm} B$$

$$\hat{R} = \sqrt{\frac{V}{W}}$$

$\hat{R}$  is an estimate of the factor by which the variance in  $\theta$  can be reduced by continuing the chains.  $\hat{R} \sim 1$  is a good sign. This should be done for all the parameters of interest.



## nested sampling



# nested sampling

Monte Carlo integration technique applied to calculating the evidence.

$$\mathcal{E} = \int d^n\theta \mathcal{L}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

- Set of  $N_a$  active points.
- Evaluate the likelihood at each of the points and sort them so that  $\mathcal{L}(\boldsymbol{\theta}_1) < \mathcal{L}(\boldsymbol{\theta}_2) < \dots < \mathcal{L}(\boldsymbol{\theta}_n)$ .
- The volume (probability) with a larger likelihood is

$$V_{pr}(\mathcal{L} > \mathcal{L}(\boldsymbol{\theta}_1)) \simeq \left(1 - \frac{1}{N_a}\right) V_o$$

where  $V_o$  is the initial volume of the parameters space.

- Pick another random point from the volume but accept it only if its likelihood is larger than minimum previously found,  $\mathcal{L}(\boldsymbol{\theta}_1)$ .

- Once we have found a good point we discard  $\theta_1$ , add the new point to the list and resort them. The new point might now be  $\theta_1$  or it might not. Now we can apply the same argument to find an estimate of the volume with  $\mathcal{L}(\theta) > \mathcal{L}(\theta_1)$ . Continue this, in the  $n$ th cycle get an estimate for the volume of

$$V_{pr}(\mathcal{L} > \mathcal{L}(\theta_1^n)) = V_{pr}^n \simeq \left(1 - \frac{1}{N_a}\right)^n V_o$$

We store all the  $\theta_1^n$ 's and  $\mathcal{L}_n \equiv \mathcal{L}(\theta_1^n)$ .

- The volume in parameter space (or probability according to the prior) associated with the likelihood  $\mathcal{L}_n$  can be found by interpolation

$$v_n = \frac{1}{2} (V_{pr}^{n-1} - V_{pr}^{n+1})$$

- Using this we can estimate the evidence as

$$\mathcal{E} \simeq \sum_{n=1}^M \mathcal{L}_n v_n$$

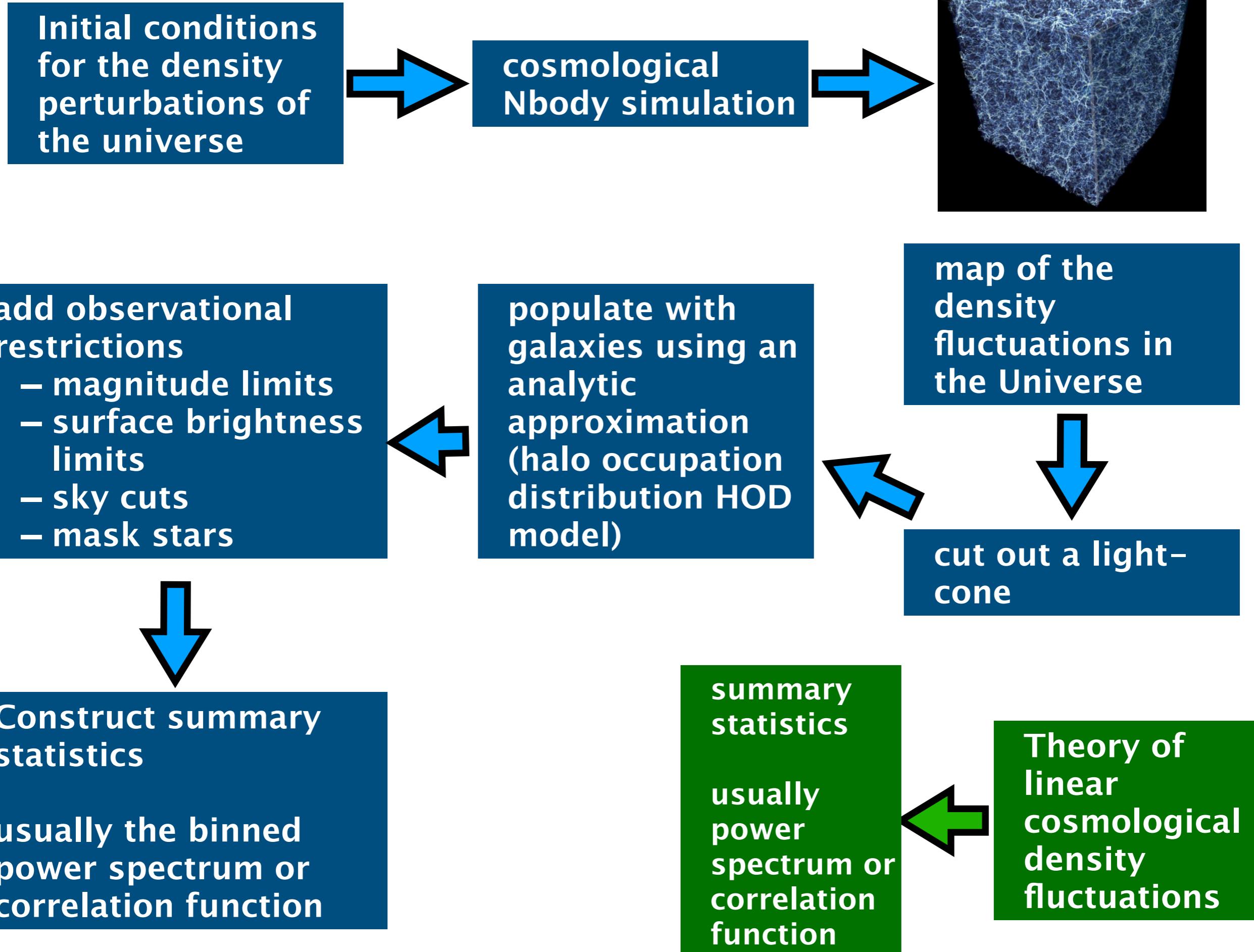
where  $M$  is the total number of cycles used.

- Any expectation value for any function of the parameters can then be estimated with

$$E [f(\boldsymbol{\theta})] \simeq \frac{\sum_{n=1}^M f(\boldsymbol{\theta}_1^n) \mathcal{L}_n v_n}{\mathcal{E}}$$

- The approximation is often made that

$$\begin{aligned} v_n &\simeq \frac{1}{2} e^{-\frac{n}{N_a}} \left( e^{+\frac{1}{N_a}} - e^{-\frac{1}{N_a}} \right) V_o \\ &\simeq \frac{e^{-\frac{n}{N_a}}}{N_a} V_o \end{aligned}$$



# Approximate Bayesian Computation (ABC)

Also called "Likelihoodless Bayesian Inference". This is for the case where the likelihood is not known in closed form, but one can simulate data sets given a set of parameters.

We would expect that if the parameters are "near" their true values the simulated data,  $D^*$  will be "near" the observed data  $D$ .

Let us invent some measure of distance between data set which we will denote  $\rho(D, D^*)$ . The number of simulations that land within  $\rho(D, D^*) < \epsilon$  would be proportional to the probability.

If the parameters are sampled according to the prior then we can make the approximation

$$p(\theta | \rho(D, D^*) < \epsilon) \sim \frac{N_\theta(\rho(D, D^*) < \epsilon)}{N(\rho(D, D^*) < \epsilon)} \sim \frac{1}{N(\rho(D, D^*) < \epsilon)} \sum_i \delta(\theta - \theta_i^*)$$

where the  $\theta_i^*$  are the parameter sets that resulted in data sets within  $\epsilon$  of the observed data.

**Initial conditions / input variables** → **simulation / theory** → **observables**

If  $\epsilon$  is very small one expects this probability to converge to the true posterior.  
Symbolically

$$p(\theta | \rho(D, D^*) < \epsilon) \xrightarrow{\epsilon \rightarrow 0} p(\theta | D) \quad (39)$$

What is  $\rho(D, D^*)$ ? The most obvious choice would be a least-squares or Euclidean distance type cost function,

$$\rho(D, D^*) = \sum_i (d_i - d_i^*)^2 \quad (40)$$

But often some statistics of the data are calculated and  $\rho(D, D^*)$  is constructed out of them.

# Variational Inference

also called :

- "Simulation-Based Inference"
- "Likelihood-Free Inference"
- "Implicit Likelihood inference"

"forward modeling" – use simulations to (partially) determine the distribution of the data

$\theta$  – parameters       $x$  – possible data sets

Joint  
probability  
is estimated

posterior

$$p(\theta, x) = p(\theta|x)p(x) \quad \text{used to estimate posterior } \{\theta_n, x_n\}$$

$$= p(x|\theta)\pi(\theta) \quad \text{used to simulate data sets}$$

prior

# Variational Inference

steps :

- simulate many data sets  $\{\theta_n, \mathbf{x}_n\}$  drawing  $\theta$  from the prior
- often the data is "compressed" to a smaller number of summary statistics
- assume a form for  $p(\theta, \mathbf{x}|\phi)$  where  $\phi$  are new "latent" parameters

This could be an analytic pdf or some other density estimation method could be used such as *normalized flow*

- maximize  $\sum_n \ln p(\theta_n, \mathbf{x}_n | \phi)$  to find  $\hat{\phi}$

This is often called minimizing the relative entropy or Kullback–Leibler divergence between the simulated data and the model distribution.

- estimate the posterior as  
 $d$  – actual data

$$p(\theta|d) = \frac{p(\theta, d|\hat{\phi})}{p(d|\hat{\phi})} = \frac{p(\theta, d|\hat{\phi})}{\int d\theta p(\theta, d|\hat{\phi})}$$



# minimum variance limit

The normalization of the likelihood

$$\int d^n x \mathcal{L}(x|\theta) = 1$$

Taking the derivative of this with respect to a parameter  $\theta_i$  gives

$$\begin{aligned} \int d^n x \frac{\partial}{\partial \theta_i} \mathcal{L}(x|\theta) &= \int d^n x \mathcal{L}(x|\theta) \frac{\partial \ln \mathcal{L}}{\partial \theta_i} \\ &= \left\langle \frac{\partial \ln \mathcal{L}}{\partial \theta_i} \right\rangle = 0 \end{aligned}$$

since  $\langle \dots \rangle = \int d^n x \mathcal{L}(x)(\dots)$ .

# The Fisher matrix

Differentiating this again gives

$$\int d^n x \left( \frac{\partial \ln \mathcal{L}}{\partial \theta_i} \frac{\partial \mathcal{L}}{\partial \theta_j} + \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} \mathcal{L} \right) = \int d^n x \left( \frac{\partial \ln \mathcal{L}}{\partial \theta_i} \frac{\partial \ln \mathcal{L}}{\partial \theta_j} \mathcal{L} + \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} \mathcal{L} \right) = 0$$

In other words

$$\begin{aligned} \mathcal{F}_{ij} &\equiv \left\langle \frac{\partial \ln \mathcal{L}}{\partial \theta_i} \frac{\partial \ln \mathcal{L}}{\partial \theta_j} \right\rangle = - \left\langle \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} \right\rangle \\ &= \text{Var} \left[ \frac{\partial \ln \mathcal{L}}{\partial \theta} \right] \quad \text{for one dimension} \end{aligned}$$

$\mathcal{F}_{ij}$  is known as the **Fisher information matrix**

Say we have an estimator for the parameter  $\theta_i$  which we will call  $\tilde{\theta}_i(\mathbf{x})$ .

$$\langle \tilde{\theta}_i \rangle = \int d^n x \tilde{\theta}_i(\mathbf{x}) \mathcal{L}(\mathbf{x}|\boldsymbol{\theta}) = \theta_i + b(\boldsymbol{\theta})$$

where  $b(\boldsymbol{\theta})$  is the bias which could be zero or not.

Taking the differential of this with respect to  $\theta_i$  gives

$$\begin{aligned} \int d^n x \tilde{\theta}_i(\mathbf{x}) \frac{\partial \mathcal{L}}{\partial \theta_i} &= 1 + \frac{\partial b}{\partial \theta_i} \\ \int d^n x \tilde{\theta}_i(\mathbf{x}) \mathcal{L} \frac{\partial \ln \mathcal{L}}{\partial \theta_i} &= 1 + b' \\ \left\langle \tilde{\theta}_i(\mathbf{x}) \frac{\partial \ln \mathcal{L}}{\partial \theta_i} \right\rangle &= 1 + b' \end{aligned}$$

It follows from (3) that

$$\left\langle (\tilde{\theta}_i(\mathbf{x}) - \theta_i) \frac{\partial \ln \mathcal{L}}{\partial \theta_i} \right\rangle = 1 + b'$$

since the extra term will be zero. This is the covariance between the estimator and the derivative of the log likelihood.

The Cauchy-Schwarz inequality applies to any covariance so

$$\left[ \left\langle (\tilde{\theta}_i(\mathbf{x}) - \theta_i) \frac{\partial \ln \mathcal{L}}{\partial \theta_i} \right\rangle \right]^2 \leq \text{Var}[\tilde{\theta}_i] \text{Var} \left[ \frac{\partial \ln \mathcal{L}}{\partial \theta_i} \right] = \text{Var}[\tilde{\theta}_i] \mathcal{F}_{ii}$$

or

$$\text{Var}[\tilde{\theta}_i] \geq \frac{(1 + b')^2}{\mathcal{F}_{ii}}$$

If the estimator is unbiased  $b' = 0$ .

This is called the **Cramér-Rao limit** or inequality. It puts a lower bound on the variance of *any* estimator.

An estimator that reaches this bound is called **efficient**. The **efficiency** of an estimator is the ratio of its variance relative to the minimum variance limit.

Directly from its definition it is easily shown that the Fisher matrix is

- 1 symmetric
- 2 the diagonal elements are  $\geq 0$
- 3 transforms like a tensor under changes of the parameters from a set  $\theta$  to  $\theta'$ ,

$$\mathcal{F}'_{ab} = \frac{\partial \theta_i}{\partial \theta'_a} \mathcal{F}_{ij} \frac{\partial \theta_j}{\partial \theta'_b}$$

# Forecasting and the Fisher matrix

Forecasting using the Fisher matrix and the Cramér-Rao limit on the variance.

$$\text{Var}[\theta] = \sigma_\theta^2 \simeq \frac{1}{\mathcal{F}_{\theta\theta}}$$

There are several criticisms of this method of forecasting errors.

Limitations:

- 1 For different fiducial parameter values the Fisher matrix can be quite different.
- 2 The Cramér-Rao limit is not likely to be reached in practice because there is no efficient estimator and/or there are unaccounted for systematic errors which dominate when the statistical errors are small.
- 3 As is, it does not account for degeneracies between parameters which might increase the marginal errors significantly. We will see that there are approximations that try to take this into account.

# Example: Simple Cosmological Supernovae

Luminosity distance:

$$\begin{aligned} D_L(z, H_o, \Omega_m, \Omega_\Lambda) &= \frac{(1+z)c}{H_o} \int_0^z dz' \frac{1}{\sqrt{\Omega_m(1+z')^3 + 1 - \Omega_m}} \\ &= \frac{c}{H_o} d_L(z, \Omega_m) \end{aligned}$$

where

$$\Omega_m + \Omega_\Lambda = 1 \quad \text{flat cosmology}$$

- $z$  is the SN's redshift
- $H_o$  is the Hubble constant
- $\Omega_m$  is the average density of the Universe in units of the critical density
- $\Omega_\Lambda$  is the cosmological constant in the same units.

Here it has been assumed that the Universe is geometrically flat. In this case the density in the cosmological constant is  $\Omega_\Lambda = 1 - \Omega_m$ .

The magnitude of the SN will be

$$\begin{aligned} m &= M_o + 5 \log_{10}(D_L(z, H_o, \Omega_m)) \\ &= M_o + 5 \log 10(H_o/c) + 5 \log_{10}(d_L(z, \Omega_m)) \end{aligned}$$

where  $M_o$  is an undetermined constant which includes the intrinsic peak luminosity.

With these assumptions the likelihood is

$$\begin{aligned} \ln \mathcal{L}(\mathbf{m}, \mathbf{z} | \Omega_m, \Omega_\Lambda) &= -\frac{1}{2} \sum_i \frac{1}{\sigma_i^2} (m_i - M_o - 5 \log 10(H_o/c) - 5 \log_{10} d_L(z_i, \Omega_m))^2 \\ &\quad - \frac{1}{2} \sum_i \ln(2\pi\sigma_i^2) \\ &= -\frac{1}{2} \sum_i \frac{1}{\sigma_i^2} (m_i - \tilde{M}_o - \mu(z_i, \Omega_m))^2 - \frac{1}{2} \sum_i \ln(2\pi\sigma_i^2) \end{aligned}$$

$M_o$  and  $H_o$  are **degenerate parameters** so they cannot be determined separately.

Now let's find the Fisher matrix

$$\frac{\partial}{\partial \Omega_m} \ln \mathcal{L}(\mathbf{m}, \mathbf{z} | \Omega_m) = - \sum_i \frac{1}{\sigma_i^2} \left( m_i - \tilde{M}_o + \mu(z_i, \Omega_m) \right) \frac{\partial \mu(z_i)}{\partial \Omega_m}$$

$$\frac{\partial^2}{\partial \Omega_m^2} \ln \mathcal{L}(\mathbf{m}, \mathbf{z} | \Omega_m) = - \sum_i \frac{1}{\sigma_i^2} \left[ \left( \frac{\partial \mu(z_i)}{\partial \Omega_m} \right)^2 + \left( m_i - \tilde{M}_o - \mu(z_i, \Omega_m) \right) \frac{\partial^2 \mu(z_i)}{\partial \Omega_m^2} \right]$$

If we take the average of this the second term will be zero because according to the likelihood  $\langle m_i \rangle = \tilde{M}_o + \mu(z_i, \Omega_m)$  so

$$\begin{aligned} \mathcal{F}_{\Omega_m \Omega_m} &= - \left\langle \frac{\partial^2}{\partial \Omega_m^2} \ln \mathcal{L}(\mathbf{m}, \mathbf{z} | \Omega_m) \right\rangle \\ &= \sum_i \frac{1}{\sigma_i^2} \left( \frac{\partial \mu(z_i)}{\partial \Omega_m} \right)^2 \end{aligned}$$

The other components of the Fisher matrix are

$$\mathcal{F}_{M_o M_o} = \sum_i \frac{1}{\sigma_i^2}$$

$$\mathcal{F}_{M_o \Omega_m} = \sum_i \frac{1}{\sigma_i^2} \frac{\partial \mu(z_i)}{\partial \Omega_m}$$

where

$$\begin{aligned} \frac{\partial \mu}{\partial \Omega_m} &= 5 \log_{10}(e) \frac{\partial}{\partial \Omega_m} \ln d_L(z) \\ &= -2.17147 \frac{(1+z)}{2d_L(z)} \int_0^z dz' \frac{(1+z')^3 - 1}{(\Omega_m(1+z')^3 + (1-\Omega_m))^{3/2}} \end{aligned}$$

We don't yet know the redshifts of the supernovae!  
 We can guess what the redshift distribution is likely to be. Let us say  
 that it is something like  $f(z) \propto x^\alpha e^{-z/z_0}$ .

$$\sum_i \rightarrow n \int dz f(z)$$

$$\mathcal{F}_{\Omega_m \Omega_m} = \frac{n}{\sigma^2} \int dz f(z) \left( \frac{\partial \mu(z)}{\partial \Omega_m} \right)^2$$

where  $f(z)$  is normalized to one and  $\sigma^2$  has been approximated as constant for all supernovae.

For 1 supernovae,  $\sigma_m = 0.3$  mag, redshift distribution parameters  $\alpha = 2$  and  $z_0 = 0.15$  the Fisher matrix is  $\mathcal{F}_{\Omega_m \Omega_m} = 1.67$ ,  $\mathcal{F}_{M_o M_o} = 11.1$ ,  $\mathcal{F}_{\Omega_m M_o} = 3.18$  for the fiducial model  $\Omega_m = 0.3$ . At a different point in parameters space,  $\Omega_m \neq 0.3$ , this will change. And for a different redshift distribution this would change.

# asymptotic normal approximations

Expand the likelihood around the MLE (or MPE for a uniform prior)

$$\begin{aligned}\ln \mathcal{L}(\mathbf{d}|\theta) &\simeq \ln \mathcal{L}(\mathbf{d}|\hat{\theta}) + (\theta - \hat{\theta}) \cdot \left. \frac{\partial \ln \mathcal{L}}{\partial \theta} \right|_{\theta=\hat{\theta}} + \frac{1}{2} (\theta - \hat{\theta})^T \left. \frac{\partial^2 \ln \mathcal{L}}{\partial \theta \partial \theta} \right|_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \mathcal{O}(|\theta - \hat{\theta}|^3) \\ &= \ln \mathcal{L}(\mathbf{d}|\hat{\theta}) + \frac{1}{2} (\theta - \hat{\theta})^T \left. \frac{\partial^2 \ln \mathcal{L}}{\partial \theta \partial \theta} \right|_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \mathcal{O}(|\theta - \hat{\theta}|^3)\end{aligned}$$

Ignoring the higher order terms, the average log-likelihood will be

$$\langle \ln \mathcal{L}(\mathbf{d}|\theta) \rangle \simeq \left\langle \ln \mathcal{L}(\mathbf{d}|\hat{\theta}) \right\rangle - \frac{1}{2} (\theta - \hat{\theta})^T \mathcal{F}(\hat{\theta})(\theta - \hat{\theta})$$

This leads us to approximate the posterior of a future experiment as

$$p(\theta) \simeq \frac{|\mathcal{F}|}{(2\pi)^{n/2}} \exp \left[ -\frac{1}{2} (\theta - \hat{\theta})^T \mathcal{F}(\hat{\theta})(\theta - \hat{\theta}) \right]$$

at least near its peak.

In this approximation:

- The *parameter* covariance matrix will be  $\mathcal{F}^{-1}$ . The variance of a single parameter *after marginalizing* over all the other parameters is

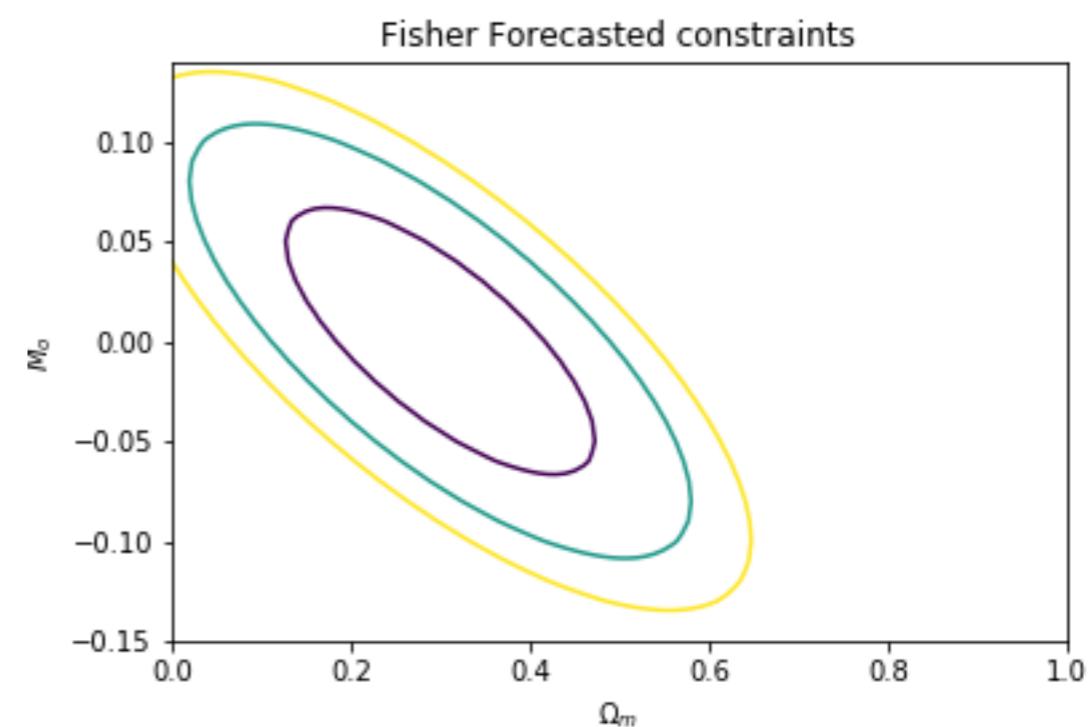
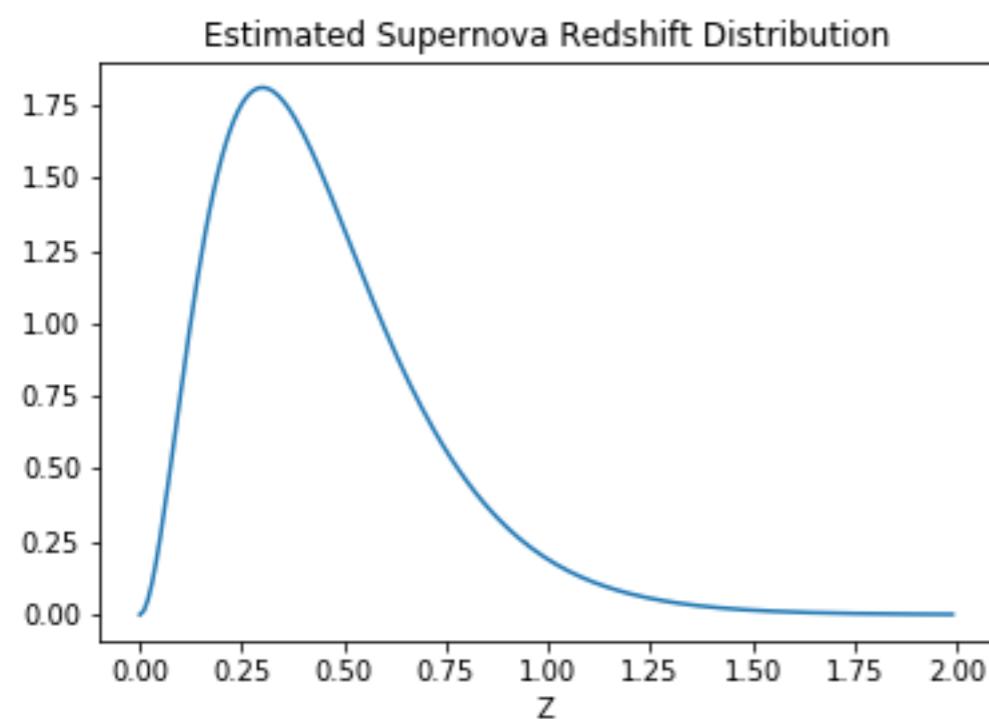
$$\sigma_\theta^2 \simeq [\mathcal{F}^{-1}]_{\theta\theta}$$

- You can easily marginalized posterior for a subsample of parameters by inverting  $\mathcal{F}$ , removing the rows and columns that correspond to the marginalized parameters and then inverting back to get  $\mathcal{F}$  in that smaller space. This comes from the rules for marginalizing a Gaussian.
- You can easily add priors on the parameters from other experiments.

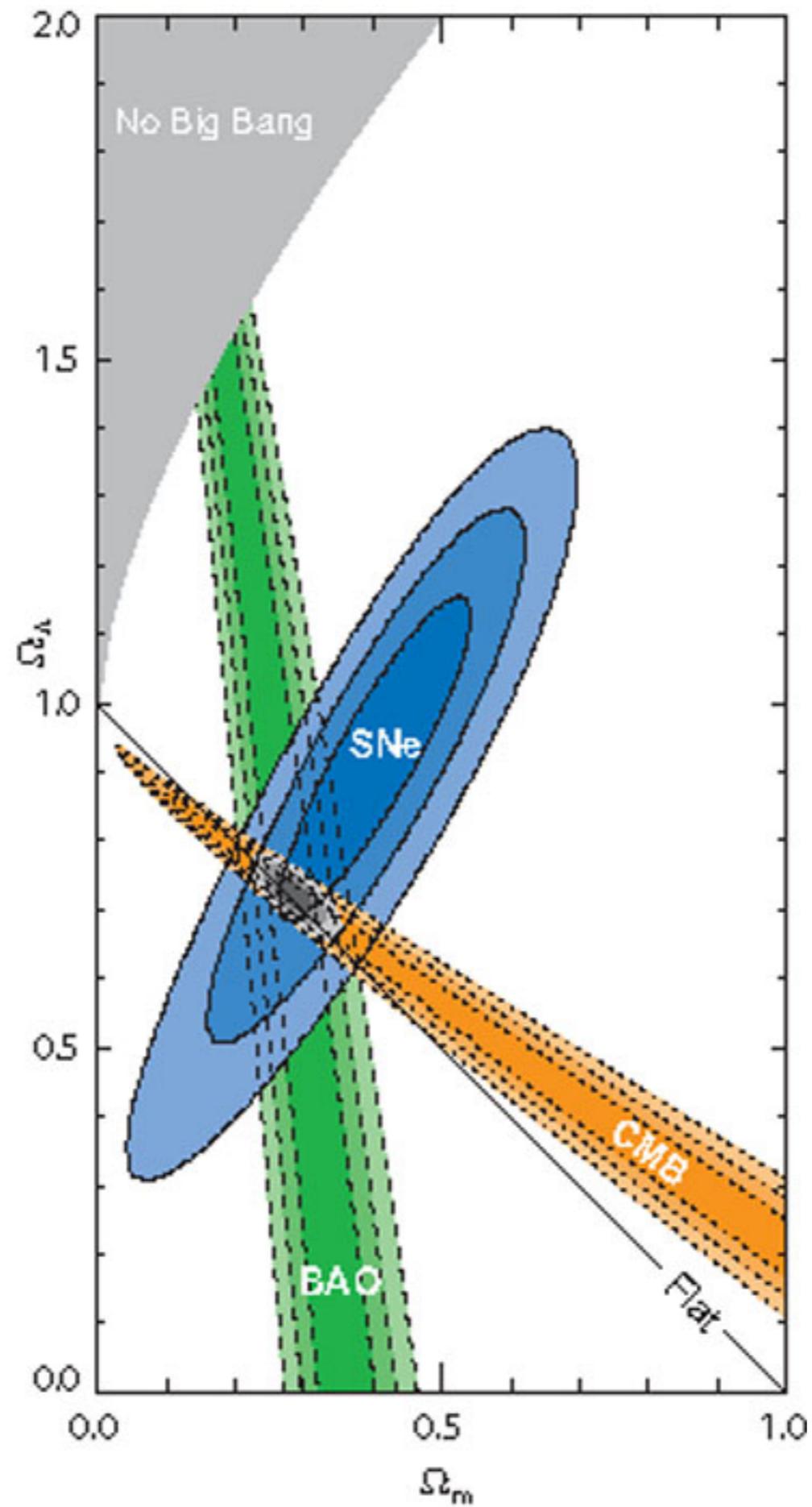
$$\mathcal{F}^{tot} = \mathcal{F} + \mathbf{C}_{\text{prior}}^{-1}$$

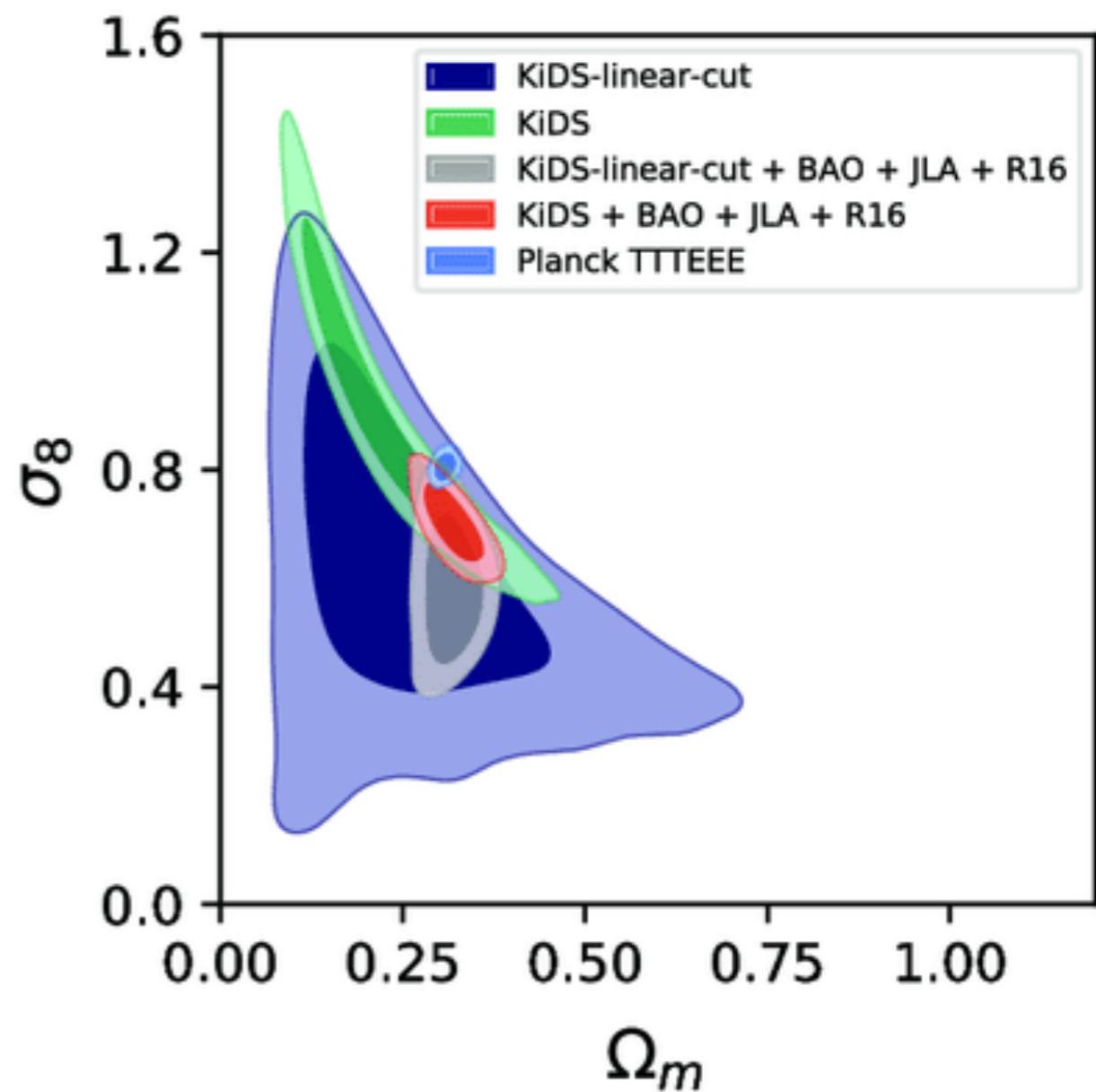
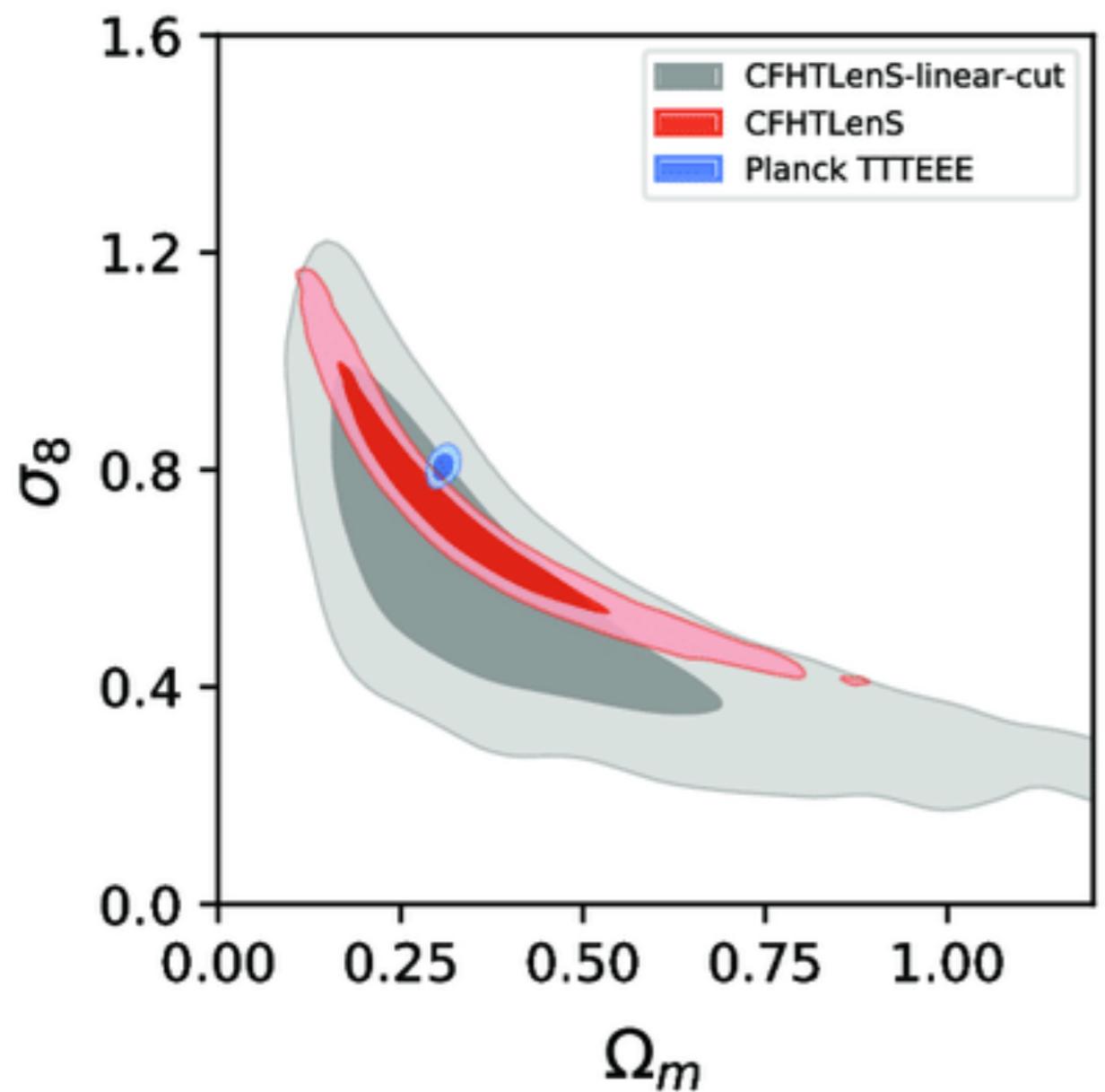
$\mathbf{C}_{\text{prior}}^{-1}$  could be the precision matrix of the parameters from some previous experiment or the Fisher matrix from some other possible experiment.

# Fisher Forecast For Hypothetical Type Ia Supernova Cosmology Survey



100 supernovae this 0.3 mag noise





# Fisher Matrix for Gaussian Distributed Data

If the data is Gaussian distributed and the mean,  $\mu$  and/or the covariance,  $C$ , of the distribution depends on some parameters  $\alpha \beta$  then the Fisher matrix takes the form

$$\mathcal{F}_{\alpha\beta} = \boldsymbol{\mu}_{,\alpha}^T \mathbf{C}^{-1} \boldsymbol{\mu}_{,\beta} + \frac{1}{2} \text{tr} [\mathbf{C}^{-1} \mathbf{C}_{,\alpha} \mathbf{C}^{-1} \mathbf{C}_{,\beta}]$$

This form of the Fisher matrix comes up a lot in Cosmology.

## Independent Gaussian distributed data

$$\mathbf{C} = \sigma^2 \mathbf{I} \quad \mathbf{C}^{-1} = \frac{1}{\sigma^2} \mathbf{I}$$

where  $\mathbf{I}$  is the identity matrix. From (38) we can find that

$$\begin{aligned}\mathcal{F}_{\mu\mu} &= \frac{n}{\sigma^2} \\ \mathcal{F}_{\mu\sigma^2} &= 0 \\ \mathcal{F}_{\sigma^2\sigma^2} &= \frac{1}{2} \text{tr} [\mathbf{C}^{-1} \mathbf{I} \mathbf{C}^{-1} \mathbf{I}] \\ &= \frac{1}{2\sigma^4} \text{tr} [\mathbf{I} \mathbf{I} \mathbf{I} \mathbf{I}] \\ &= \frac{n}{2\sigma^4}\end{aligned}$$

So there is no unbiased estimator for the mean that have a variance smaller than  $\sigma^2/n$ . The sample mean is unbiased and has a variance of  $\sigma^2/n$  so it is an *efficient* estimator.

# Asymptotic behavior of the maximum likelihood estimator

The log likelihood of  $n$  independent data points, or data sets,  $x_n$  can be written

$$\ln(\mathcal{L}(x|\theta)) = \ln\left(\prod_i^n L(x_i|\theta)\right) = \sum_i^n \ln(L(x_i|\theta))$$

If  $\hat{\theta}(x)$  is the MLE and  $\theta_o$  is the true value of the parameter, under some often satisfied requirements on the regularity of the likelihood function, in the limit of large amounts of independent data:

- 1  $\hat{\theta}$  converges to  $\theta_o$  in probability. This means that as the amount of data gets very large the MLE will eventually get arbitrarily close to the true value. In other words the MLE is a *consistent* statistic.
- 2 The MLE is asymptotically normally distributed

$$(\hat{\theta} - \theta_o) \sim \mathcal{N}\left(0, \frac{1}{nF_{\theta\theta}}\right)$$

where  $F_{\theta\theta}$  is the Fisher matrix for  $L(x|\theta_o)$ .



# logistic regression

We are interested in predicting some outcome  $Y$  that can be true or false.

We wish to be able to predict the probability of  $Y$  given independent variables  $X$ .

Each trial has a different probability of being correct,  $p_i$ , depending on the variable  $x_i$ . For the two possibilities  $y_i = 0$  or  $1$  the likelihood is binomial

$$\mathcal{L}(\{y_i\}) = \prod_{i=0}^n p_i^{y_i} (1 - p_i)^{1-y_i} \quad (8)$$

The idea behind logistic regression is to make a model for the  $p_i$ 's that depends on the  $x_i$ 's and some parameters. The best fit parameters can be found by maximizing the likelihood.

The most popular model is a linear one for the log of the odds  $p_i/(1 - p_i)$

$$\ln \left( \frac{p_i}{1 - p_i} \right) = \theta_o + \sum_j \theta_j x_j \quad (9)$$

$$= \theta_o + \boldsymbol{\theta} \cdot \mathbf{x} \quad (10)$$

The free parameters are the  $\theta_j$ 's.

The odds implies that the probability is in the form of a **sigmoid function**,  $S(u)$ ,

$$S(u) = \frac{1}{1 + e^{-u}} = \frac{e^u}{1 + e^u} \quad (11)$$

The probability that the  $i$ th data point will be true is given by

$$p_i = S(\theta_o + \boldsymbol{\theta} \cdot \mathbf{x}) \quad (12)$$

This is plugged into the likelihood 8 for the training set. The maximum likelihood or maximum relative entropy solution for  $\boldsymbol{\theta}$  must be found by numerically

*entropy*

$$S[p] = - \sum_i p_i \log p_i$$

*relative entropy*

*or*

*Kullback-Leibler  
divergence*

$$D_{KP}[q|p] = \sum_i q_i \log \left( \frac{q_i}{p_i} \right)$$

*cross entropy*

$$H[q|p] = - \sum_i q_i \log (p_i)$$

$$D_{KL}[q|p] = H[q|p] - S(q)$$

*Maximizing the likelihood is the same as :*

*Minimizing the relative entropy or Kullback-Leibler divergence  $D_{KL}$  between the bootstrap distribution of the data and the model distribution*

*Minimizing the cross-entropy between the two.*

# multinomial logistic regression

What is there are more than 2 classes? Let's say there are  $K$  classes. The probability of being in class  $i$  is modeled as

$$p_i(\mathbf{x}) = \frac{1}{1 + \sum_{k \neq p} e^{\boldsymbol{\theta}^k \cdot \mathbf{x}}} \times \begin{cases} e^{\boldsymbol{\theta}^k \cdot \mathbf{x}} & , \quad k \neq p \\ 1 & , \quad k = p \end{cases} \quad (13)$$

where  $k = p$  is the "pivot class". This function, called a **softmax** function.

There are now  $(K - 1) \times N_f$  coefficients  $\boldsymbol{\theta}^k$  where  $N_f$  is the number of features.

There are several strategies for finding the coefficients.

- maximize the multinomial likelihood with respect to these coefficients by numerical means.

- Another is to find the coefficients  $\boldsymbol{\theta}^k$  by doing a binomial logistic fit between class  $k$  and the pivot class  $p$ . This is repeated for all  $K - 1$  classes besides  $p$ .



# Information & Entropy

Shannon's Axioms:

- Axiom I :  $S$  is a real continuous function of the probabilities  $p_i$ ,  $S[p_1, \dots, p_m]$ .
- Axiom II : If all  $p_i$ 's are equal,  $p_i = 1/m$ , then  $S[1/m, 1/m, \dots, 1/m]$  is an increasing function of  $m$ . If all states are equally probable increasing the number of states increases the uncertainty.
- Axiom III: The grouping property. For all possible inclusive groupings  $g = 1 \dots N$  of the states  $i = 1 \dots n$  we must have

$$S = S[\{P\}] + \sum_g P_g S_g$$

where

$$P_g = \sum_{i \in g} p_i$$

The unique functional that satisfies these axioms is the **Shannon entropy or information**

$$S[p] = - \sum_i^m p_i \ln p_i$$

Some properties:

- $S[p] \geq 0$
- For the uniform probability case  $p_i = 1/m$ , the maximum ignorance case,  $S[1/m, \dots, 1/m] = \ln(m)$ .
- In the case of complete certainty one of the probabilities will be 1 and all the others 0. In this case  $S[1, 0, \dots, 0] = 0$ .
- If there are two variable  $x$  and  $y$  with joint probability  $p(x, y)$  their entropy is

$$S_{xy} = - \sum_{xy} p(x, y) \ln p(x, y)$$

and if they are independent  $p(x, y) = p(x)p(y)$  so

$$S_{xy} = S_x + S_y \quad \text{independent variable}$$

and in general

$$S_{xy} \leq S_x + S_y$$

You can define the analogous entropy of a continuous distribution,

$$S[p] = - \int_{-\infty}^{\infty} dx \ p(x) \ln(p(x))$$

This entropy has the important flaw that it is not coordinate invariant.

# the maximum entropy principle for choosing a distribution

The maximum entropy principle, sometime abbreviated MaxEnt, holds that **the best distribution to use for a variable is the one that has the maximum entropy (or least information) subject to any prior constraints on the distribution**. The idea is that you should assume the least possible information beyond your constraints and this is quantified by the entropy, i.e. maximum ignorance.

The maximum entropy distribution with only the normalization constraint ( $\sum_i^m p_i = 1$ ) is of course  $p_i = 1/m$  and the maximum entropy is  $S_{max} = \ln(m)$ . This is the state of maximal ignorance or minimum information.

Now let us say that we have a constraint on the variance of a continuous distribution, namely

$$\text{Var}_p[x] = \sigma^2$$

where  $\sigma^2$  is a constant.

We can find the MaxEnt distribution using Lagrangian multipliers

$$\delta \left\{ - \int_{-\infty}^{\infty} dx \ p(x) \ln p(x) - \lambda_o \left( \int_{-\infty}^{\infty} dx \ p(x) - 1 \right) - \lambda_1 \left( \int_{-\infty}^{\infty} dx \ x^2 p(x) - \sigma^2 \right) \right\} = 0$$

giving

$$-\ln p(x) - 1 - \lambda_o - \lambda_1 x^2 = 0$$

or

$$\begin{aligned} p(x) &= \exp(-1 - \lambda_o - \lambda_1 x^2) \\ &= A e^{-\lambda_1 x^2} \end{aligned}$$

where  $A$  is a normalization constant.

$A$  and  $\lambda_1$  are determined by the two constraints so

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

The distribution with the maximum degree of ignorance subject to the constraint on the variance is the normal distribution.

This is another reason to favor the normal distribution that has no apparent relation to the central limit theorem.

The entropy of a normal distribution is

$$S_{norm} = \frac{1}{2} + \frac{1}{2} \ln (2\pi\sigma^2) .$$

# Connection to Statistical Physics

**Boltzman:** The number of ways to occupy  $m$  states with  $N$  particles

$$\Omega = \frac{N!}{n_1! \dots n_m!}$$

with  $\sum_i^m n_i = N$ . Using Sterling's approximation

$$\begin{aligned}\ln \Omega &= N \ln N - N - \sum_i^m (n_i \ln n_i - n_i) \\ &= N \ln N - \sum_i^m n_i \ln n_i \\ &= N \left[ \ln N - \sum_i^m \frac{n_i}{N} \left( \ln \frac{n_i}{N} + \ln N \right) \right] \\ &= -N \sum_i^m \frac{n_i}{N} \ln \left( \frac{n_i}{N} \right) \\ &= -N \sum_i^m p_i \ln p_i\end{aligned}$$

if  $p_i = n_i/N$ . You can see the clear resemblance to Shannon's entropy.

**Gibbs' entropy** Gibbs changed the interpretation of  $p_i$  to be the probability of a collective state  $i$  where the particles (spins, molecular species, etc.) are not necessarily independent and not the single particle occupation of single particle states.

Consider a series of constraints on the expectation values of the form

$$\langle f^k \rangle = \sum_i^m f_i^k p_i = F^k$$

For example,  $F^k$  could be the average energy  $\bar{E}$  in which case  $f_i = \epsilon_i$  the energy of a specific state  $i$ . Or  $f_i^k$  is the number of particles or molecules of a certain type in state  $i$  then  $F^k$  will be the average number of those species. Or  $F^k$  could be the magnetic polarization, etc. The state of the macroscopic system is labeled by the  $F^k$ 's.

The canonical distribution is found by maximizing the entropy

$$-\sum_i^m p_i \ln p_i - \lambda_o \left( \sum_i^m p_i - 1 \right) - \lambda_k \left( \sum_i^m f_i^k p_i - F^k \right) = \text{const.}$$

$$-\ln p_i - 1 - \lambda_o - \lambda_k f_i^k = 0$$

or

$$p_i = e^{-\lambda_o - 1} e^{-\lambda_k f_i^k}$$

$$p_i = \frac{e^{-\lambda_k f_i^k}}{Z} \quad Z = \sum_i e^{-\lambda_k f_i^k} = e^{\lambda_o - 1}$$

You can recognize  $Z$  as the partition function.

and see that

$$\frac{\partial \ln Z}{\partial \lambda_k} = -F^k$$

The maximum entropy is

$$\begin{aligned} S_{max} &= - \sum_i p_i \ln p_i \\ &= - \sum_i \frac{e^{-\lambda_k f_i^k}}{Z} (-\lambda_k F^k - \ln Z) \\ &= \lambda_k F^k + \ln Z \end{aligned}$$

$$\begin{aligned} \frac{\partial S_{max}}{\partial F^k} &= \frac{\partial \lambda_i}{\partial F^k} F^i + \lambda_k + \frac{\partial \ln Z}{\partial \lambda_k} \frac{\partial \lambda_i}{\partial F^k} \\ &= \frac{\partial \lambda_i}{\partial F^k} F^i + \lambda_k - F^k \frac{\partial \lambda_i}{\partial F^k} \\ &= \lambda_k \end{aligned}$$

For the canonical ensemble  $F^k$  is the average energy of the system,  $\bar{E}$ ,  $f_i^k$  is the energy of state  $i$ ,  $\epsilon_i$  and

$$\frac{\partial S_{max}}{\partial \bar{E}} \equiv \frac{1}{k_B T} = \lambda_\epsilon$$

so the Lagrangian multiplier associated with the internal energy is the inverse of the temperature. In this case equation (30) becomes the perhaps familiar

$$S_{max} = \frac{\bar{E}}{k_b T} + \ln Z$$

The energy  $\bar{E}$  (and the  $\epsilon_i$ 's) will also depend on the volume

$$\frac{\partial S_{max}}{\partial V} = \frac{\partial S_{max}}{\partial \bar{E}} \frac{\partial \bar{E}}{\partial V} = -\frac{P}{k_B T}$$

Likewise, the Lagrangian multiplier associated with the average number of a chemical species is the the chemical potential,  $\mu_k/k_B T$ .

From (26), the canonical distribution is then

$$p_i \propto \exp \left[ -\frac{\epsilon_i}{k_b T} \right]$$

The thermodynamic "entropy of the system" is actually the maximum of the entropies that are consistent with the constraints. From an information theory prospective, the thermodynamic entropy is the negative the amount of information that is needed to specify a microstate given a specified macrostate which corresponds to a fixed average energy, volume and number of particles.

# Maximum entropy for Inference

There exists a functional  $S[p, q]$  that ranks all possible posteriors  $p(x)$  in order of preference given the prior information  $q(x)$ .

Caticha's Axioms:

- Axiom 1 : *Locality* In the absence of information about some domain  $D$  the probability should not change,  $p(x|D) = q(x|D)$ .
- Axiom 2: *Coordinate invariance*  $S[p, q]$  should remain the same when the coordinates are changed.
- Axiom 3: *Consistency for independent systems subsystems* When a system is composed of subsystems that are known to be independent, it should not matter whether the inference procedure treats them separately or jointly.

# Maximum entropy for Inference

Amazingly, just these three axioms lead to a unique functional,

$$S[p, q] = - \int dx \ p(x) \ln \left( \frac{p(x)}{q(x)} \right)$$

The **relative entropy**.

## Principle of Maximum Entropy Updating:

*Given prior knowledge  $q(x)$  and some new information, the best posterior  $p(x)$  is the one that maximizes the relative entropy while being consistent with the new information.*

# relative entropy

The relative entropy is also called the **Kullback–Leibler divergence** or distance.

An important property is

$$S(p|q) \geq 0$$

and  $S(p|q) = 0$  only when  $p(x) = q(x)$ . As required by the axioms,  $S(p|q)$  is invariant under transformations of the random variables  $x$ .

The relative entropy is often used as a measure of how distant two distribution are from each other. It is not a true distance however because it is not symmetric,  $S[p, q] \neq S[q, p]$

$S[p, q]$  can be interpreted as the amount of information gained when the distribution is updated from  $q(x)$  to  $p(x)$ . For example, the information a new experiment adds to our knowledge of some parameters can be quantified in this way. This can be useful in planning an experiment when many parameters are being measured.

Another situation in which this comes up is when a new posterior is found, perhaps in a high dimensional parameter space, and one wants to quantify how much extra information has been gained by including the latest data. The relative entropy of the posterior with respect to the prior can be used to quantify this.

For reference and some intuition the relative entropy of two Gaussians is

$$S[p, q] = \frac{1}{2} \left[ \ln \left( \frac{|\mathbf{C}_q|}{|\mathbf{C}_p|} \right) + \text{tr} \left[ \mathbf{C}_p (\mathbf{C}_q^{-1} - \mathbf{C}_p^{-1}) \right] + (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^T \mathbf{C}_q^{-1} (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q) \right]$$

In univariate case this is

$$\begin{aligned} S[p, q] &= \frac{1}{2} \left[ \ln \left( \frac{\sigma_q^2}{\sigma_p^2} \right) + \frac{\sigma_p^2}{\sigma_q^2} - 1 + \frac{(\mu_p - \mu_q)^2}{\sigma_q^2} \right] \\ &= \frac{1}{2} \left[ \ln \left( \frac{\sigma_q^2}{\sigma_p^2} \right) + \frac{(\sigma_p^2 - \sigma_q^2) + (\mu_p - \mu_q)^2}{\sigma_q^2} \right] \end{aligned}$$

The first part expresses a change in the variances or constraining power of the distributions and the second comes from a mismatch in the means of the distributions.

The relative entropy of the posterior to the prior is sometimes called the **surprise**

$$\begin{aligned}
 S = S[p(\boldsymbol{\theta}|D), \pi(\boldsymbol{\theta})] &= \int_{-\infty}^{\infty} d\boldsymbol{\theta} \ p(\boldsymbol{\theta}|D) \ln \left[ \frac{p(\boldsymbol{\theta}|D)}{\pi(\boldsymbol{\theta})} \right] \\
 &= \int d\boldsymbol{\theta} \ p(\boldsymbol{\theta}|D) \ln \left[ \frac{\mathcal{L}(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\mathcal{E}(D)\pi(\boldsymbol{\theta})} \right] \\
 &= \langle \ln [\mathcal{L}(D|\boldsymbol{\theta})] \rangle_{p(\boldsymbol{\theta}|D)} - \ln \mathcal{E}(D)
 \end{aligned}$$

This term is also used for other measures of how different the posterior is from the prior. This is a measure of the information gain that comes from the data.

In the special case where the likelihood is Gaussian, the prior is uniform over a volume  $V_\pi$  and the likelihood is very small on all the borders of this prior volume so that integrals over the posterior are not effected by it, the surprise reduces to

$$S = -\frac{d}{2} (1 + \ln(2\pi)) - \frac{1}{2} \ln |\mathbf{C}| + \ln V_\pi$$

while the evidence is

$$\ln \mathcal{E} = \ln \mathcal{L}^{max} + \frac{d}{2} \ln(2\pi) + \frac{1}{2} \ln |\mathbf{C}| - \ln V_\pi$$

# equivalence of maximum likelihood distribution & minimum relative entropy

Let us say we have an experiment that has a categorical outcome. There are  $k$  possible outcomes. The number of observed events in category  $i$  is  $n_i$ . We have a model that predicts the probability of outcome  $i$  as  $p_i(\theta)$ . The parameters of this model are  $\theta$ .

The likelihood is a multinomial distribution

$$\mathcal{L}(\{n_i\}|\theta) = \frac{N!}{\prod_i^k n_i!} \prod_i^k p_i(\theta)^{n_i}$$

We can find the maximum likelihood by taking the derivative

$$\begin{aligned}\frac{\partial}{\partial \theta} \ln \mathcal{L}(\{n_i\}|\theta) &= \frac{\partial}{\partial \theta} \left[ \ln N! + \sum_i^k n_i \ln p_i(\theta) - \sum_i^k n_i! \right] \\ &= \sum_i^k n_i \frac{\partial}{\partial \theta} \ln p_i(\theta) = 0\end{aligned}$$

Now let's look at this problem differently. The empirical distribution is  $p_i = n_i/N$ . We could look for the distribution within the family of distributions parameterized by  $\theta$  that is "closest" to the empirical distribution. To define "closest" we might use the minimum relative entropy or Kullback–Leibler distance between the empirical and trial distributions.

$$S \left[ \frac{n_i}{N}, p_i(\theta) \right] = - \sum_i^k \frac{n_i}{N} \ln \left( \frac{n_i}{N p_i(\theta)} \right)$$

Its minimum occurs at

$$\frac{\partial}{\partial \theta} S = \frac{1}{N} \sum_i^k n_i \frac{\partial}{\partial \theta} \ln p_i(\theta) = 0$$

which is clearly the same solution as for the maximum likelihood.

In machine learning the relative entropy is often used as a cost function. In this context one has a training set  $\{\mathbf{y}, \mathbf{x}\}$  where  $\mathbf{x}$  are the feature or independent variables and  $\mathbf{y}$  are the dependent variables. The model gives a probability  $p_i(\theta, \mathbf{x})$ . The relative entropy is minimized to find the best  $\theta$ . We can see now that this is equivalent to finding the maximum likelihood solution.

I have presented this as for categorical dependent variables, but you can see that it works fine for the continuous variables as well. In this case all the  $n_i$ 's are one.

