TIME-DOMAIN ASTRONOMY

Lectures 2: Introduction to Statistics and Probability

Stefano Covino
INAF / Brera Astronomical Observatory

Probability

- We adopt here a Bayesian view: probability is just a degree of certainty about a statement.

- An **experiment** is any action that can have a set of possible results where the actually occurring result cannot be predicted with certainty prior to the action.

- The set Ω of all outcomes of an experiment is known as the **outcome space** or sample space.

- A well-balanced coin toss gives Ω={H,T}, and the inherent symmetries of the experiment leads to P(H)=P(T)=0.5

Axioms of probability

- A probability space consists of the triplet {Ω, F, P}, a sample space, a class of events, and a function that assigns a probability to each event in F following:

Axiom 1 $0 \leq P(A) \leq 1$, for all events A

Axiom 2 $P(\Omega) = 1$

Axiom 3 For mutually exclusive (pairwise disjoint) events $A_1, A_2, \ldots,$

$$P(A_1 \cup A_2 \cup A_3 \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots,$$

that is, if for all $i \neq j, A_i \cap A_j = \emptyset$ ($\emptyset$ denotes the empty set or null event), then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

- From these, it is easy to show, for generic events C and D, that:

$$P(C \cup D) = P(C) + P(D) - P(C \cap D);$$

Conditional Probabilities

- A far from trivial concept in probability:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

- The formula above can be read as “probability of A given B (i.e. knowing that B occurred)”.

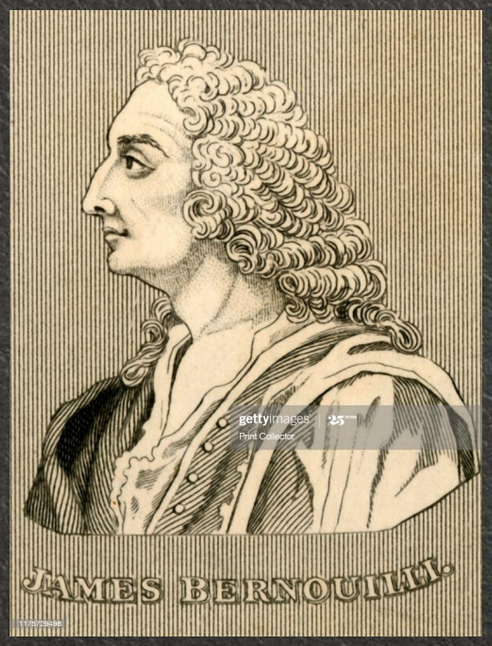
- E.g., rolling a dice, and with A={1,2,3}, implies P(A)=1/2. The probability of an even outcome (B={2,4,6}) is again 1/2, but P(A ∩ B) is 1/6. Then, P(A|B) = 1/3.

Bayes' Theorem

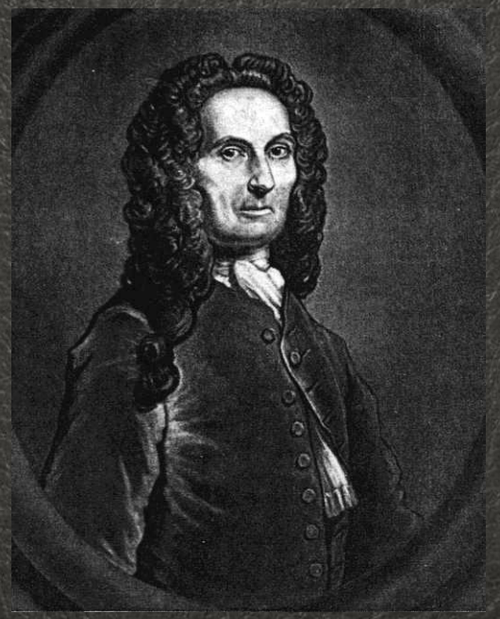
- Due to Thomas Bayes, but recognized earlier by James Bernoulli and Abraham de Moivre, and later fully explicated by Pierre Simon de Laplace.

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

Thomas Bayes



James Bernouille



Abraham de Moivre

Pierre Simon de Laplace

Bayes' Theorem

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

- P(A|B): the **posterior**, or the probability of the model parameters given the data: this is the result we want to compute.
- P(B|A): the **likelihood**, the same function, but with a different meaning, used in the frequentist approach.
- P(A): the **model prior**, which encodes what we knew about the model prior to the application of the data.
- P(B): the **data probability** or **evidence**, in most cases just a normalization factor.

It is intriguing that most of the debate between "frequentists" vs "Bayesians" is not due to the mathematics of the theorem but to its philosophical meaning, i.e. the basis of Bayesian inference.

Basic definitions and some useful "frequentist" concepts

- *Random variable*: variable describing the possible outcome of an experiment. Can be either:
  - *Continuous* (e.g. measurement of the temperature at noon in Utrecht)
  - *Discrete* (e.g. counting the number of passengers on board Bus 12)
- *Parent distribution*: distribution of values for RV if experiment is repeated an infinite number of times.

Mean of parent distribution:

$$\mu \equiv \lim_{N \to \infty} \left(\frac{1}{N} \sum_i x_i \right)$$

where x_i are N measurements

Median of parent distribution:

50% of x_i values < $\mu_{1/2}$ and 50% of x_i values > $u_{1/2}$
To calculate $u_{1/2}$ from experimental data, x_i must be sorted - computationally inconvenient.

Mode or most probable value:

The most likely value to occur, i.e.

$$P(\mu_{\max}) \geq P(x \neq \mu_{\max})$$

Mean, Median and Mode identical for symmetric distr.

Standard deviation (σ) and variance (σ^2) of *parent distribution*:

$$\sigma^2 \equiv \lim_{N\to\infty}\left[\frac{1}{N}\sum_i (x_i-\mu)^2\right]$$

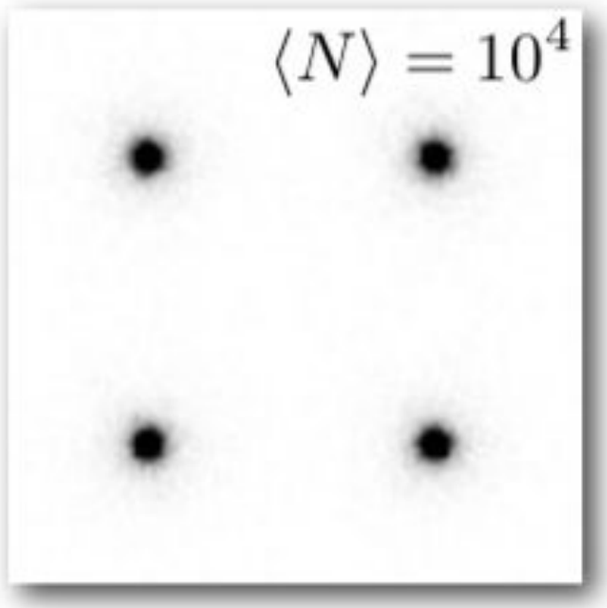
Standard deviation (s) and variance (s^2) of a *sample population*:

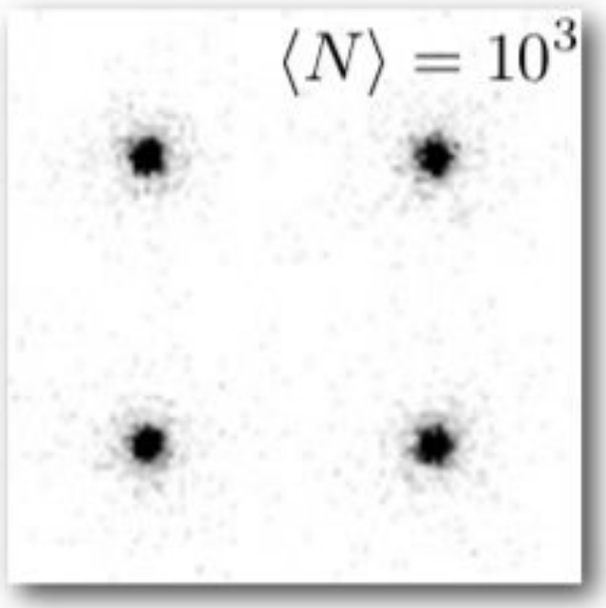
$$s^2 = \frac{1}{N-1}\sum_i (x_i-\overline{x})^2$$

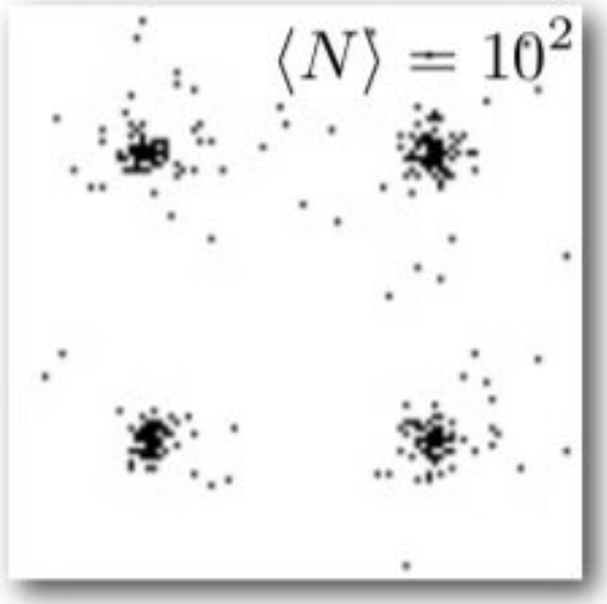
Note factor N-1 instead of N. This is to account for the fact that the sample mean has been determined from the *data*.

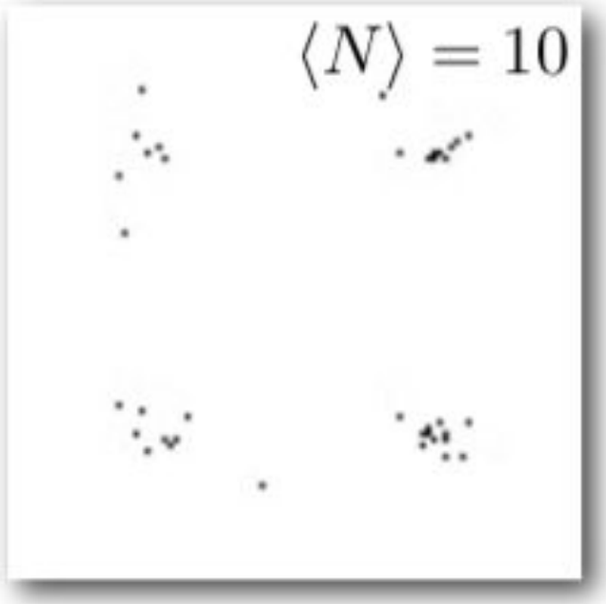
Counting statistics

- Most astronomical observations involve detection of electromagnetic radiation
- This is made up of discrete quanta - photons
- Many modern detectors count photons more or less directly
- Any number of counts *N* has a random uncertainty associated with it.









Photon counting is a *random process*.

Repeated measurements will not yield identical results but be subject to random fluctuations.

How to quantify this?

Counting statistics

Assume mean number of detections μ in some time interval *t*.

Divide *t* into *n* sub-intervals, where *n*>>μ. Then the probability of detecting a photon in one sub-interval is *p*~μ/*n*.

Probability of detecting *k* photons during *t* is then (binomial distr.):

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

$$P(k;\mu) = \frac{n!}{k!(n-k)!} \left(\frac{\mu}{n}\right)^k \left(1-\frac{\mu}{n}\right)^{n-k}$$

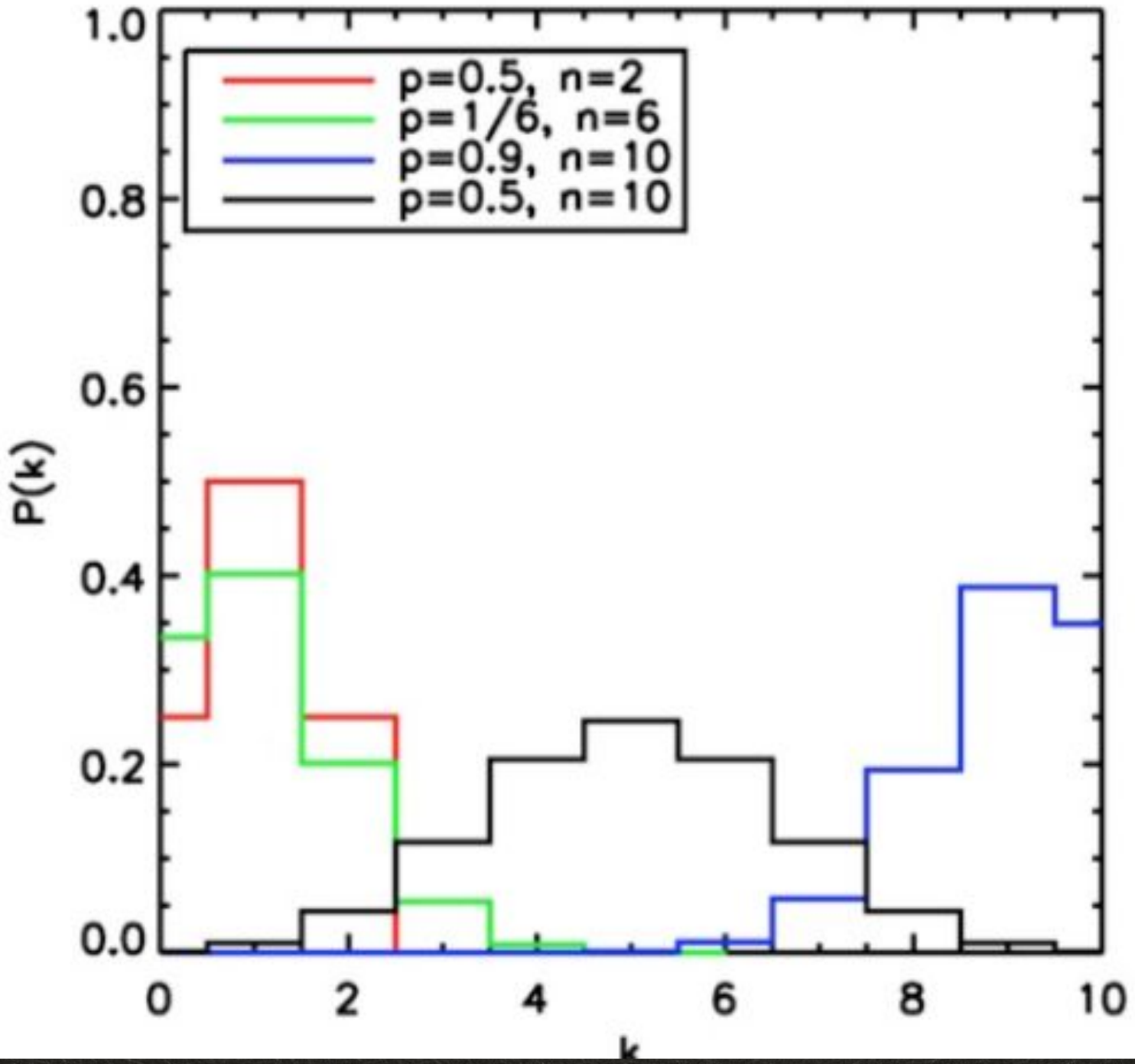
which in the limit of large n reduces to the *Poisson* distribution,

$$P(k;\mu) = \frac{\mu^k}{k!} e^{-\mu}$$

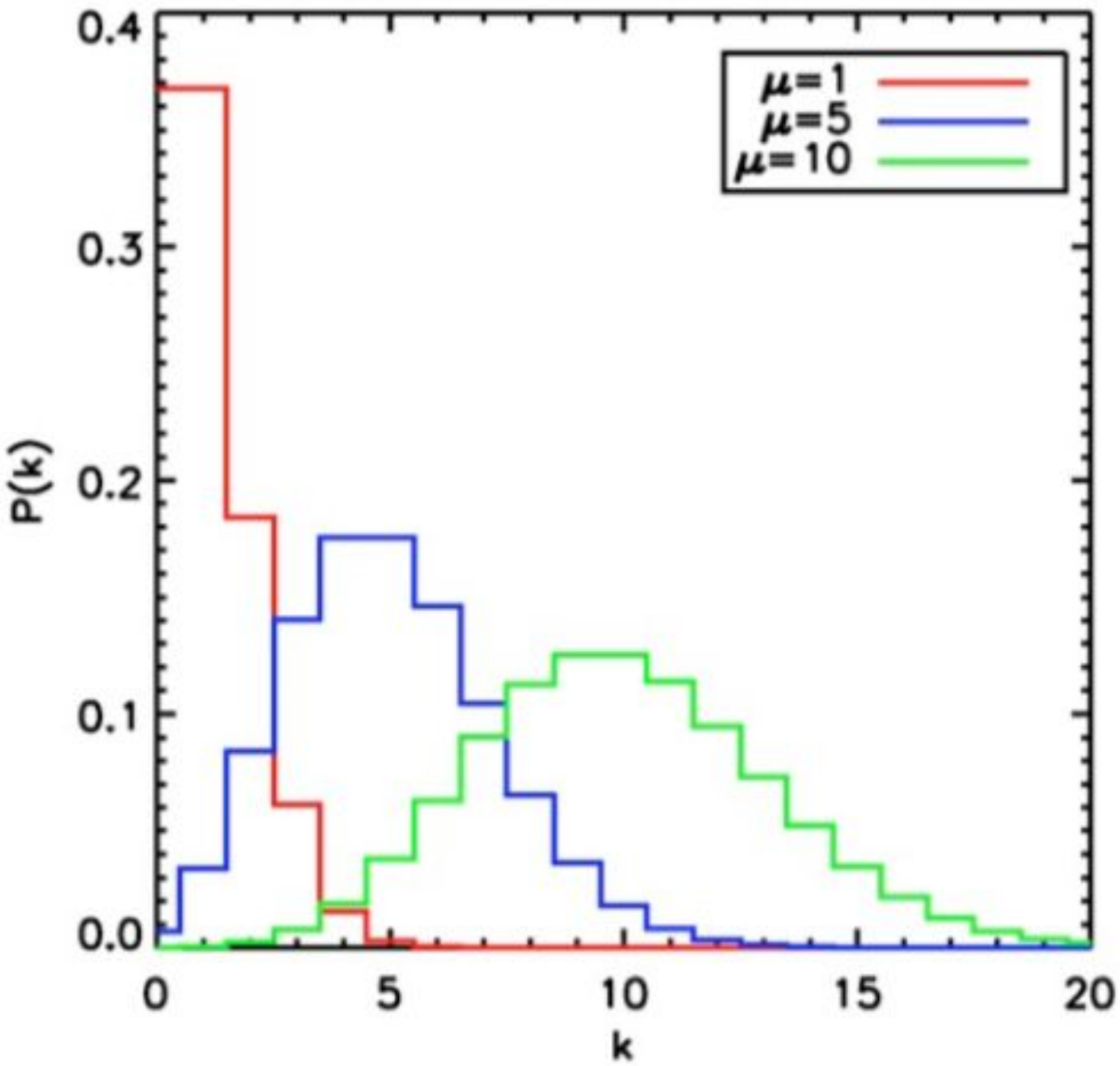
Binomial / Poisson

- Binomial distribution:
  - Applicable when carrying out a well-defined number of trials *n*, each of which has a probability *p* of success. B.D. gives the probability of having exactly *k* successful outcomes
  - Example: Rolling n=100 dice, what is the probability of getting exactly k=15 sixes (p=1/6)? $P_{binom}(15, 100, 1/6) = 10\%$
- Poisson distribution:
  - Applicable when the mean number μ of successful outcomes is known (or can be estimated) and is much smaller than the maximum possible number of successful outcomes (e.g. $p<<1$)
  - Example: About 1% of pregnancies are twin pregnancies. In 1000 pregnancies ($\mu=10$), what is the probability of having exactly 5 twin pregnancies? $P_{poisson}(5, 10) \sim 4\%$

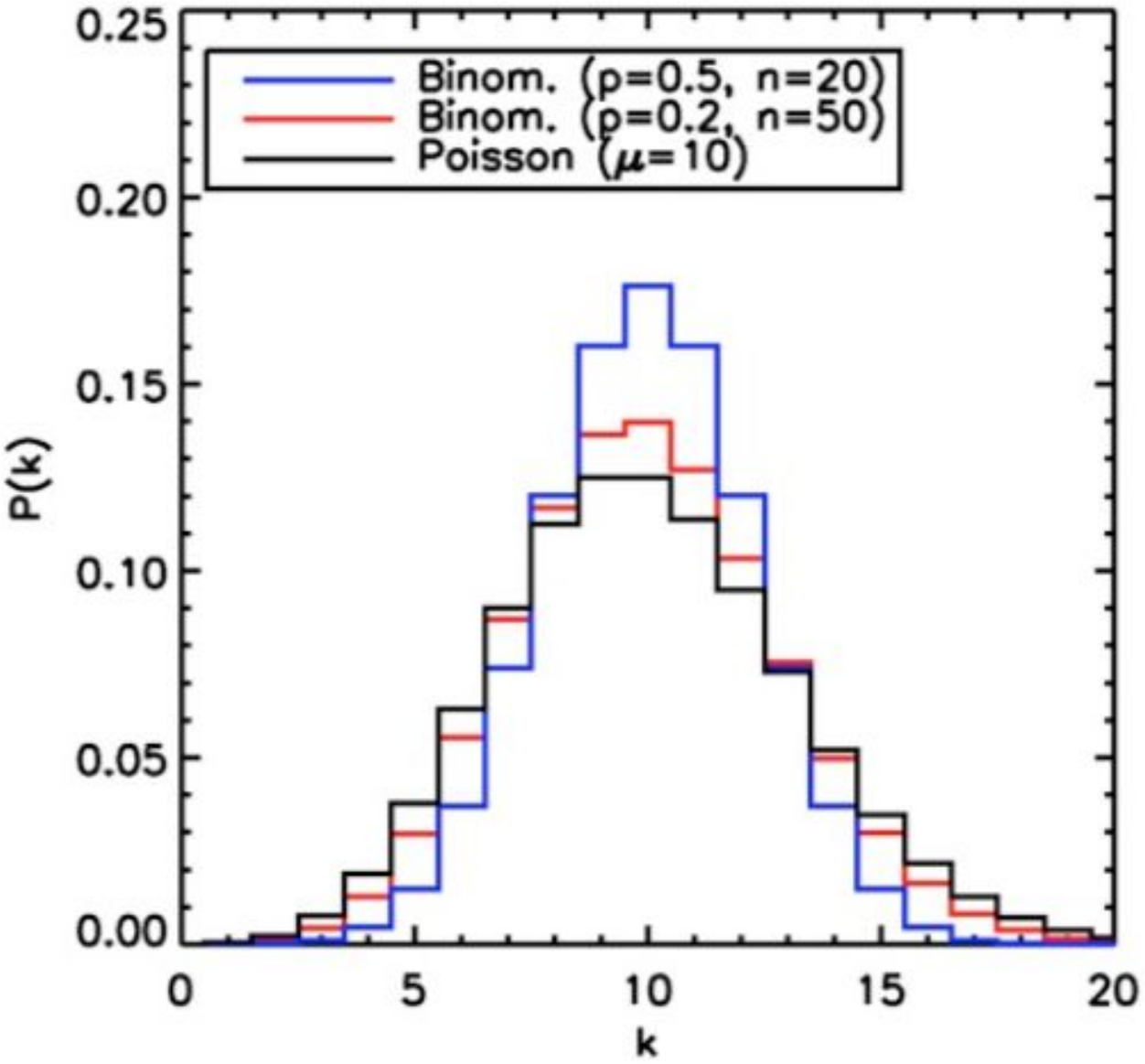
Binomial distributions



Poisson distributions



Binomial / Poisson



"Tail" of Poisson distr. extends to infinity (in principle)

"Tail" of binomial distr. only extends to *n*.

Mean and Variance

For any probability density function of a discrete variable $P(x_i)$:

Mean: $\mu = \sum_{i=0}^{\infty} x_i P(x_i)$ Variance: $\sigma^2 = \sum_{i=0}^{\infty} (x_i - \mu)^2 P(x_i)$

For binomial distribution:

$$\mu = \sum_{0}^{n} kP(k;n,p) = \sum_{0}^{n} k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = np$$

$$\sigma^2 = \sum_{0}^{n} (k-np)^2 \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = np(1-p)$$

In the limit of large n and small p (i.e. *Poisson* distribution)

$$\sigma^2_{\text{Poisson}} = np = \mu$$

Error on the Mean

Mean of N measurements $x_1, x_2, ..x_n$:

$$\overline{x} = \frac{1}{N}\sum_i x_i$$

Propagation of error on mean:

$$\sigma_{\overline{x}}^2 = \left(\frac{\partial \overline{x}}{\partial x_1}\right)^2 \sigma_{x_1}^2 + \left(\frac{\partial \overline{x}}{\partial x_2}\right)^2 \sigma_{x_2}^2 + \cdots$$

$$= \left(\frac{1}{N}\right)^2 \sigma_{x_1}^2 + \left(\frac{1}{N}\right)^2 \sigma_{x_2}^2 + \cdots \qquad = \frac{1}{N^2}\sum_i \sigma_{x_i}^2$$

If $\sigma_{x1}=\sigma_{x2} .. = \sigma_i$,

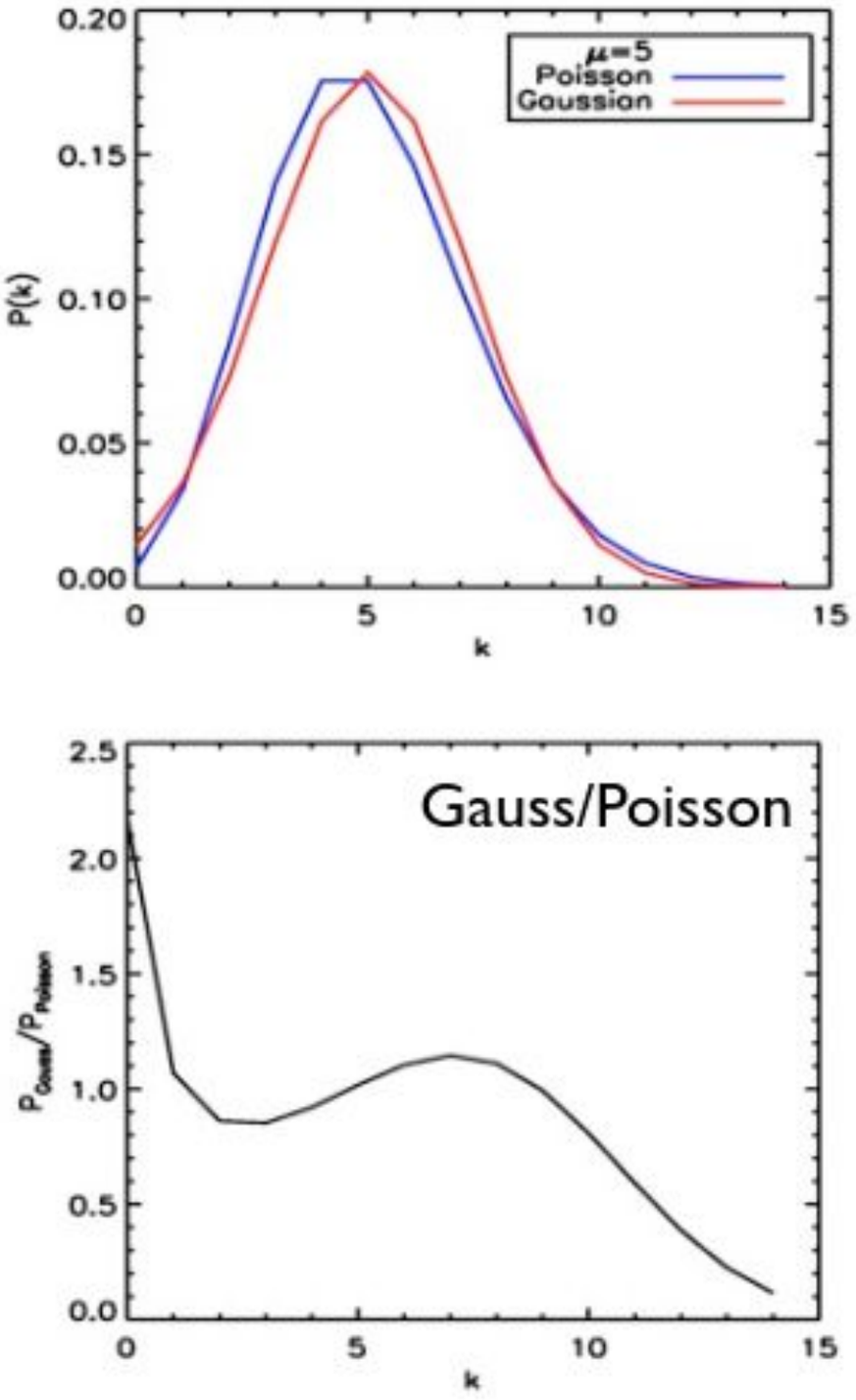
$$\sigma_{\overline{x}}^2 = \frac{N\sigma_i^2}{N^2} = \frac{\sigma_i^2}{N} \qquad \text{i.e.} \qquad \boxed{\sigma_{\overline{x}} = \frac{\sigma_i}{\sqrt{N}}}$$

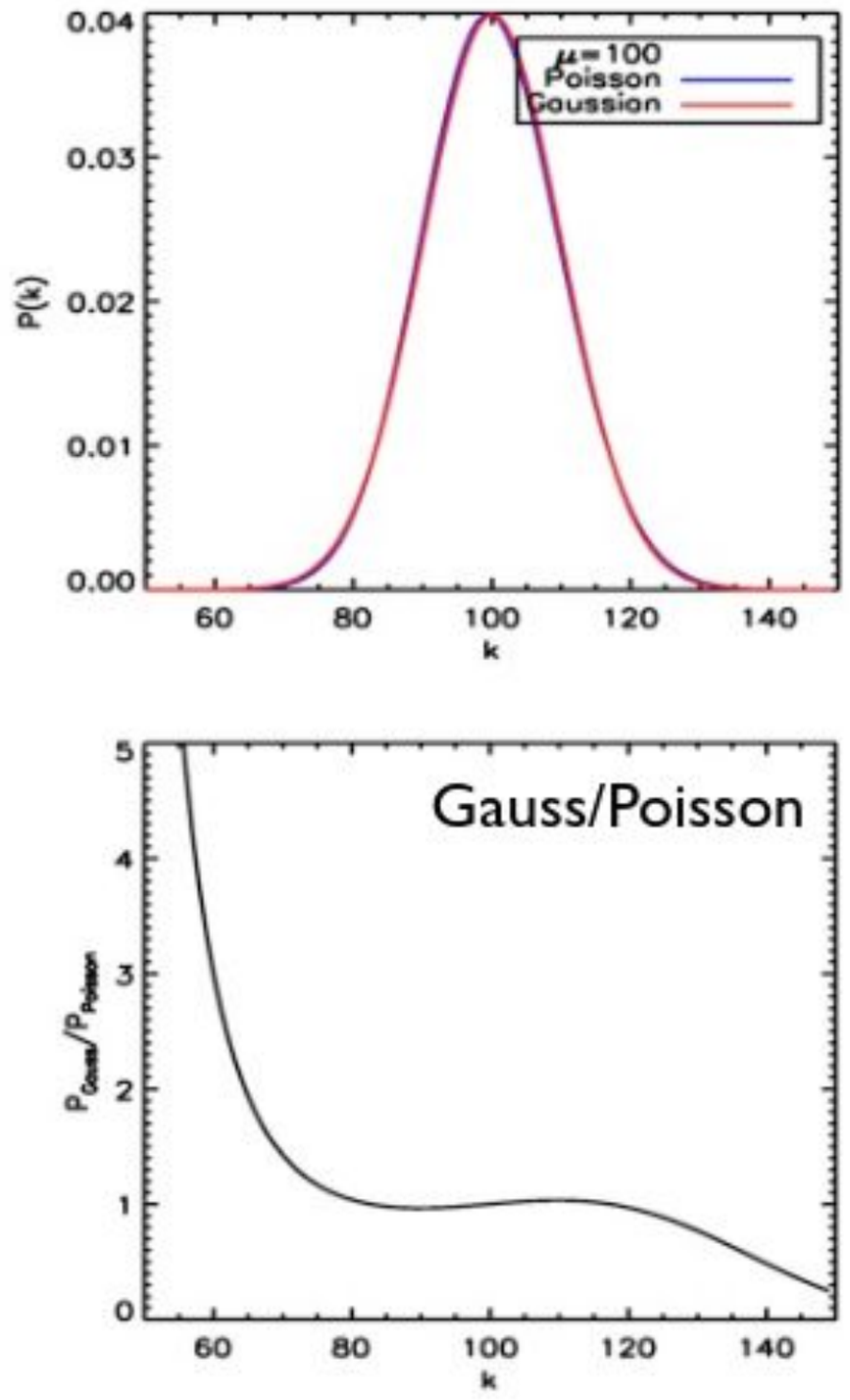
The Normal Distribution

For large μ, the Poisson distribution can be approximated by a *Gaussian*:

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

Much easier to compute; does not involve factorials, but *only an approximation to the Poisson distribution.*


μ=5
Poisson
Gaussian
P(k)
k
0.20
0.15
0.10
0.05
0.00
0
5
10
15


μ=100
Poisson
Gaussian
P(k)
k
0.04
0.03
0.02
0.01
0.00
60
80
100
120
140

Gauss/Poisson
k
2.5
2.0
1.5
1.0
0.5
0.0
0
5
10
15

Gauss/Poisson
k
5
4
3
2
1
0
60
80
100
120
140

Gaussian errors II: Central Limit Theorem

"The convolution of a large number of (positive) functions $f_i(x)$ with variances σ_i^2 is a Gaussian with variance $\sigma^2 = \Sigma\sigma_i^2$ "

Note that the individual f_i do not have to be Gaussian!

Convolution

- Mathematically, the convolution of two functions f(x) and g(x) defined as:

$$(f * g)(x) \equiv \int_{-\infty}^{\infty} f(\xi) g(x - \xi) d\xi$$

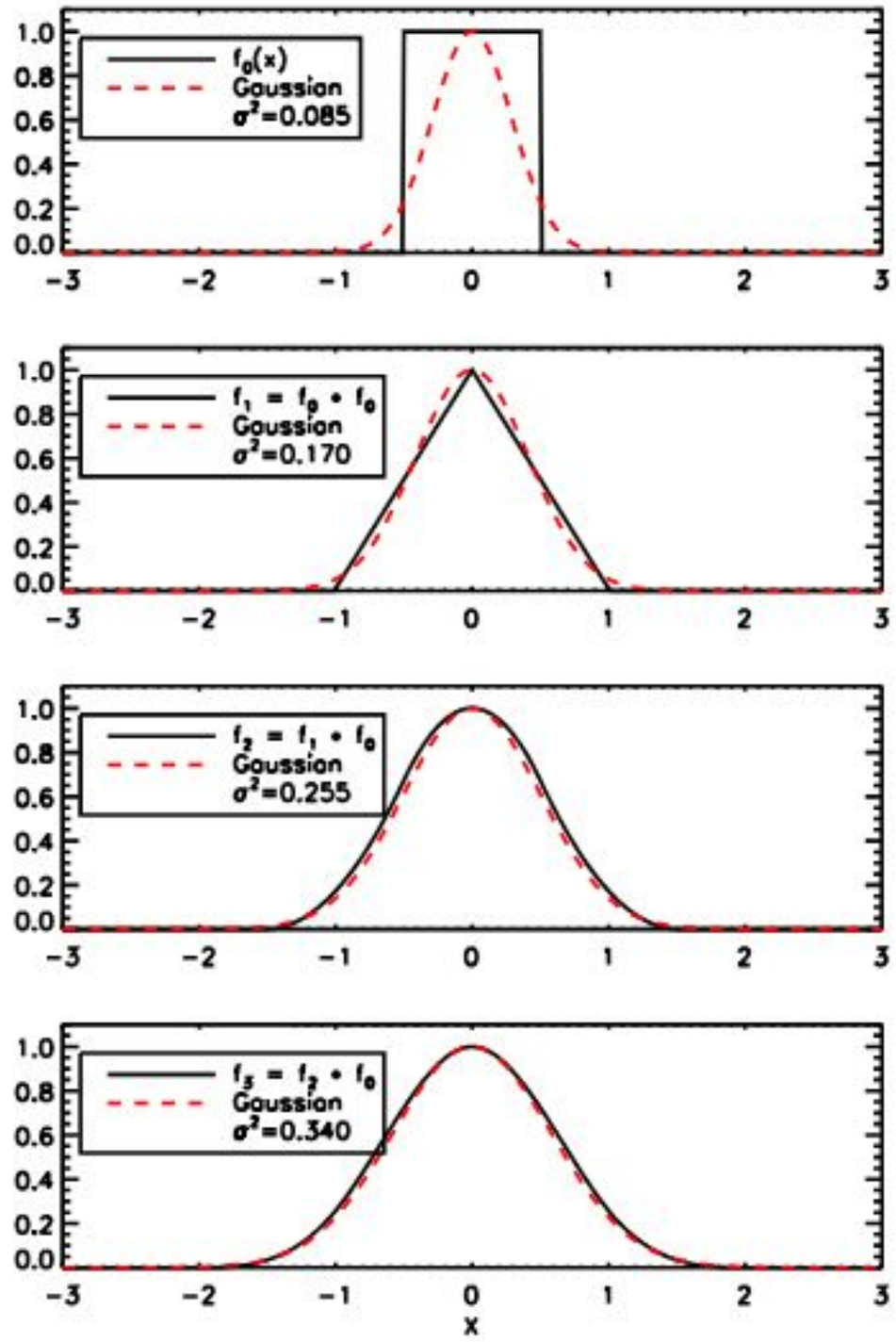
- Conveniently carried out using Fourier transforms:

$$\mathcal{F}\{f * g\} = \mathcal{F}\{f\} \cdot \mathcal{F}\{g\}$$

Central Limit Theorem

"The convolution of a large number of (positive) functions $f_i(x)$ with variances σ_i^2 is a Gaussian with variance $\sigma^2 = \Sigma\sigma_i^2$ "

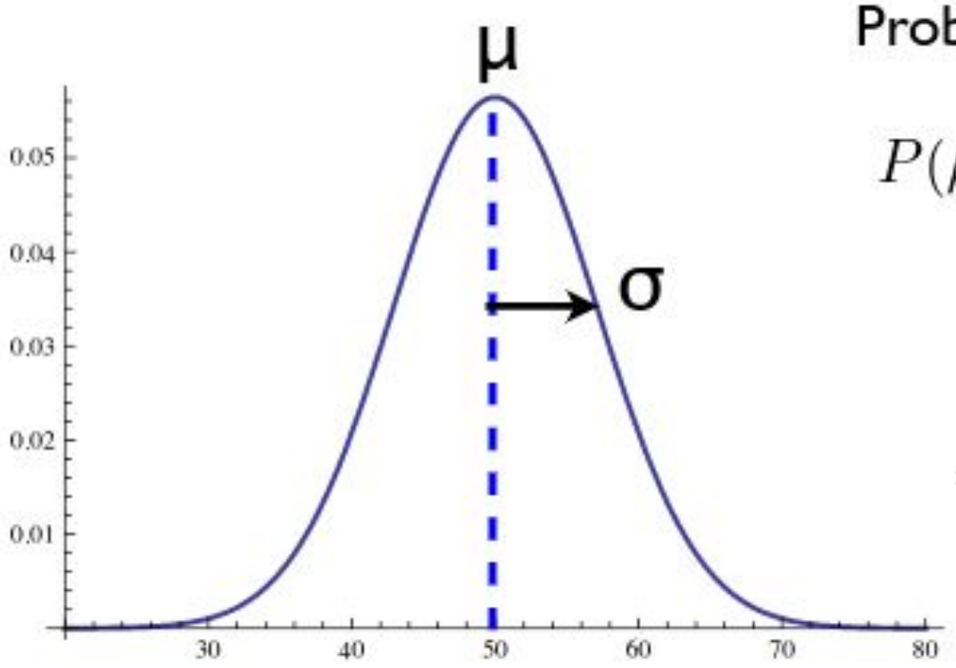
Note that the individual f_i do not have to be Gaussian!



Central limit theorem II

- Whenever a measurement follows from a combination of several sub-measurements, each with its own (possibly non-normal) error distribution, the result is likely to have normally distributed errors.
- Example: a CCD chip is "read out" through a number of electronics components, each of which introduces some noise. The result is Gaussian "read-out noise".

Standard deviation and confidence intervals



Probability of outcome within 1σ of μ:

$$P(\mu-\sigma \le x \le \mu+\sigma) = \int_{\mu-\sigma}^{\mu+\sigma} P(x)dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu-\sigma}^{\mu+\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

$$= 68.2\%$$

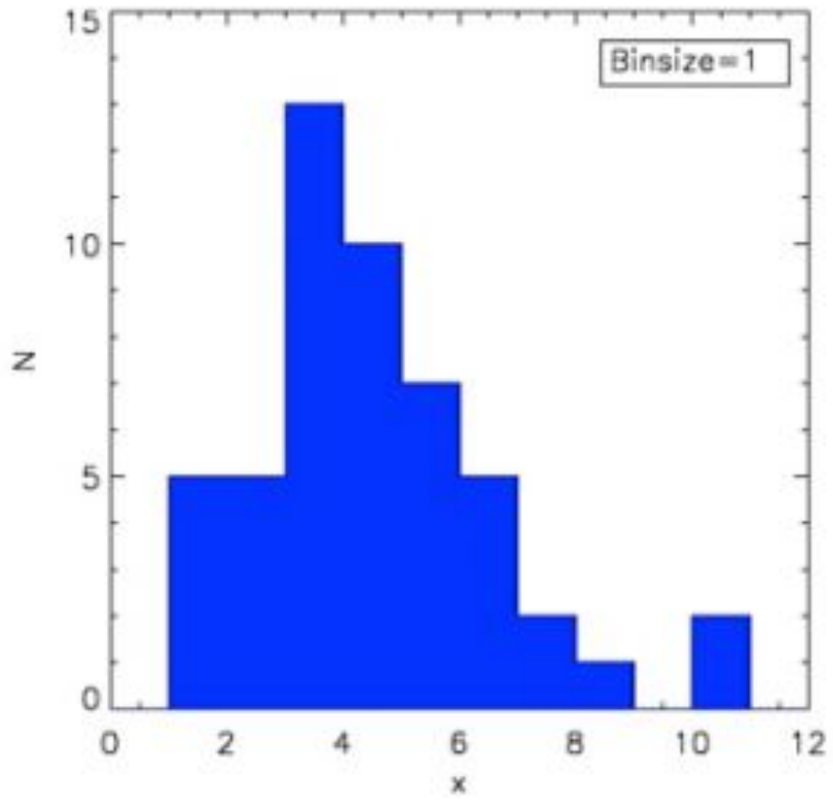
n	1	2	3	5
$P(\mu-n\sigma<x<\mu+n\sigma)$	68.2%	95.5%	99.7%	$1-5.7\times10^{-7}$

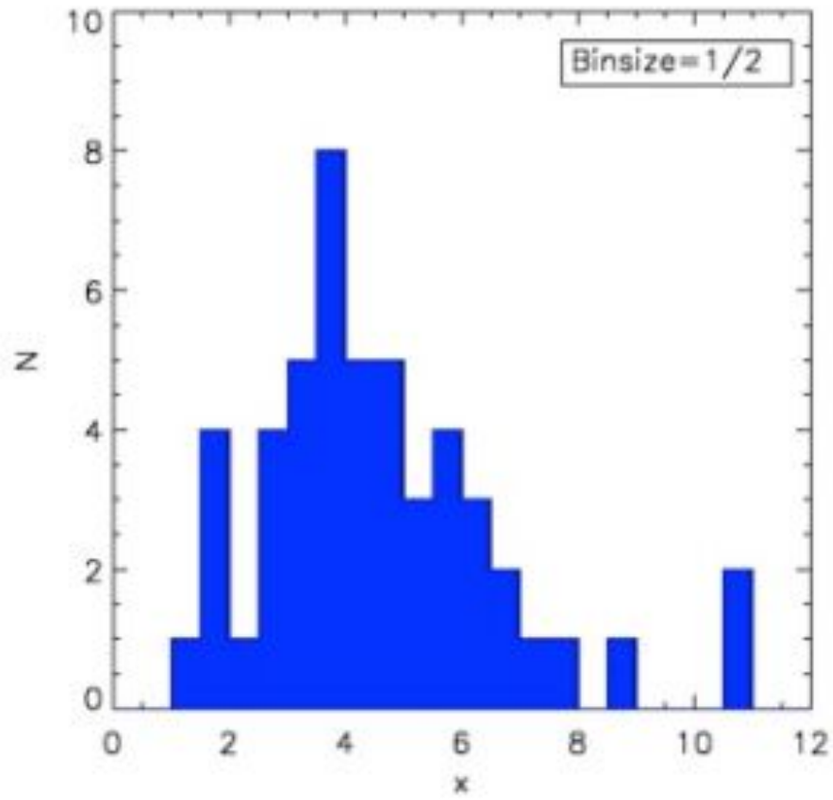
Parameter estimation

Given a set of n measurements $x_1, x_2, .. x_n$:

1.71395,	3.99812,	3.59233,	3.68776,	5.38543,	4.49257,
3.16461,	3.71537,	2.68601,	5.03422,	3.51358,	3.86742,
3.01930,	7.14149,	5.82127,	10.6515,	5.79230,	3.92880,
5.94576,	3.19683,	6.64978,	4.47039,	1.44092,	6.24392,
3.95624,	10.7719,	4.31418,	4.50722,	1.71201,	3.24773,
1.95836,	2.69437,	6.80226,	1.51801,	4.95176,	2.58177,
2.70607,	5.49929,	4.91509,	7.60173,	4.03818,	8.58724,
4.55077,	3.11701,	5.90127,	4.45284,	2.46260,	6.19397,
4.86509,	6.08224				

E.g. assuming that the parent distribution is a Gaussian, how do we estimate μ and σ?





We could fit a Gaussian directly to a histogram of the data.
But is this the best way?

- Which binsize is the best to use?
- How to weigh different bins?
- What to do about empty bins?
- etc., etc.

Maximum likelihood

- Maximum likelihood estimation: For a given set of observations x_i, and an assumed parent distribution $P(x; \xi_1, \xi_2, ..)$ find the values of $\xi_1, \xi_2, ...$ for which the probability of observing the values x_i is maximised.

Maximum likelihood

Probability ("likelihood") of observed x_i for a particular μ,σ, assuming Gaussian PDF:

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

Total *likelihood* of observing the whole dataset:

$$P_G(\mu, \sigma) = \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left[-\frac{1}{2}\sum_i \left(\frac{x_i - \mu}{\sigma}\right)^2\right]$$

Maximised when $\frac{\partial P_G(\mu, \sigma)}{\partial \mu} = 0$ and $\frac{\partial P_G(\mu, \sigma)}{\partial \sigma} = 0$

$$\frac{\partial}{\partial \mu}\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left[-\frac{1}{2}\sum_i\left(\frac{x_i-\mu}{\sigma}\right)^2\right]=0$$

$$\frac{\partial}{\partial \mu}\exp\left[-\frac{1}{2}\sum_i\left(\frac{x_i-\mu}{\sigma}\right)^2\right]=0$$

.. and after a bit of manipulation:

$$\mu=\frac{1}{N}\sum_i x_i=\overline{x}$$

I.e. just the usual definition of sample mean!

Maximum likelihood

- For a normal (Gaussian) distribution we recover the (intuitively reasonable) result that the most likely value of the mean of the parent distribution is just the sample mean.
- However, M-L may be employed to estimate the parameters of *any* parent distribution that best describe a given dataset (e.g. the slope of a power-law)
- Avoids *binning* of data.

Maximum likelihood

- Analytic solution not always possible - then the likelihood function $P_{ML}(\xi_1, \xi_2, ..)$ may need to be minimised numerically.
- Computation of likelihood function can often lead to numerical underflow - convenient to operate on its logarithm:

$$\log P_{\mathrm{ML}}(\xi_1, \xi_2, \ldots) = \log \prod_i P(x_i; \xi_1, \xi_2, \ldots) = \sum_i \log P(x_i; \xi_1, \xi_2, \ldots)$$

What about errors?

- In many cases (as in the previous example) we do not have an analytic expression for the desired parameter.
- Sometimes it is impractical or impossible to estimate errors by direct propagation
- Make use of *Monte-Carlo* methods

Bootstrap estimation of variance:

For function $T(x_i)$ of n measurements $\{x_i\}$, define m random subsamples $\{T_j\}$ by drawing n data from $\{x_i\}$ at random *with replacement*.

If T is the parameter value for the full sample, the variance can be estimated as

$$\sigma_T^2 = \frac{1}{m}\sum_{j=1}^{m}(T_j - T)^2$$

Example:

$\{x\} = \{16, 11, 15, 16, 11, 12, 18, 5, 3, 1\}$ $\mu=10.8$, $\mu_{1/2} = 12$

20 random sub-samples with replacement:

$\{x_1\} = \{11, 11, 3, 3, 15, 12, 1, 3, 5, 11\}$ $\mu_1=7.5$, $\mu_{1/2,1} = 11$

$\{x_2\} = \{16, 12, 15, 5, 11, 18, 11, 16, 15, 15\}$ $\mu_2=13.4$, $\mu_{1/2,2} = 15$

.....

$\{x_{20}\} = \{5, 15, 11, 11, 12, 3, 12, 15, 5, 3\}$ $\mu_{20}=9.2$, $\mu_{1/2,2} = 11$

$$\sigma_\mu = \sqrt{\frac{1}{20}\sum_{i=1}^{20}(\mu - \mu_i)^2} = 1.69$$

$$\sigma_{\mu_{1/2}} = \sqrt{\frac{1}{20}\sum_{i=1}^{20}(\mu_{1/2} - \mu_{1/2,i})^2} = 2.13$$

Note: $\sigma_{\mu_{1/2}} > \sigma_\mu$

Median vs Mean

- For symmetric parent distributions (e.g. Gaussian), mean and median give same result in the limit of infinite number of measurements
- Median less sensitive to "outliers"
- However, median is less *precise* than mean: generally, $\sigma_{1/2} \sim 4/3\ \sigma$
- Example: Combining N images. Using median instead of mean will decrease S/N by ~25%. Since $S/N \sim N^{1/2}$, median combination is equivalent to using only $(3/4)^2 \sim 0.56$ N images!

Hypothesis testing

- How can we be sure that a given model applied to data is the correct one?
- Example 1: we can fit a straight line ($y = ax + b$) to any set of measurements (x_i, y_i) but are the data *consistent* with a linear relation?
- Example 2: we can calculate the sample mean and variance for any set of data, but are the data actually consistent with a Gaussian parent distribution?

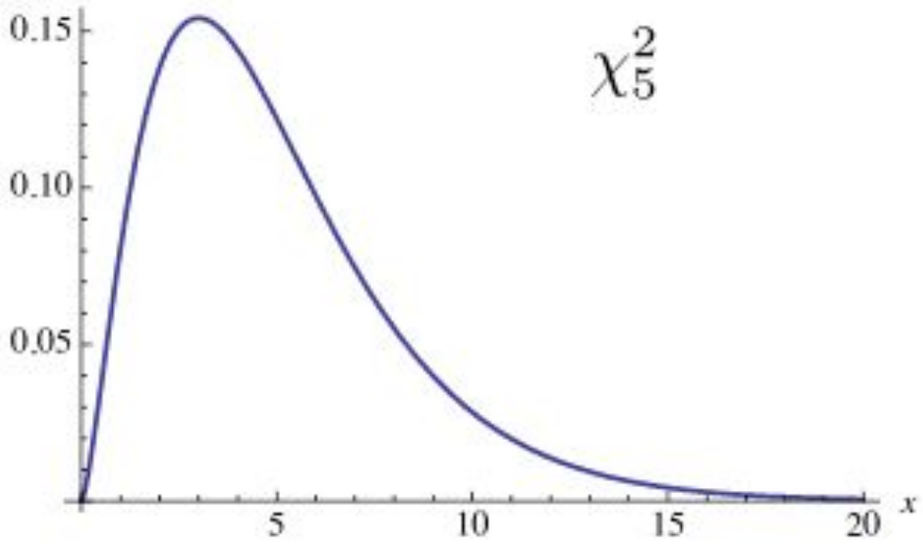
The χ^2 statistics

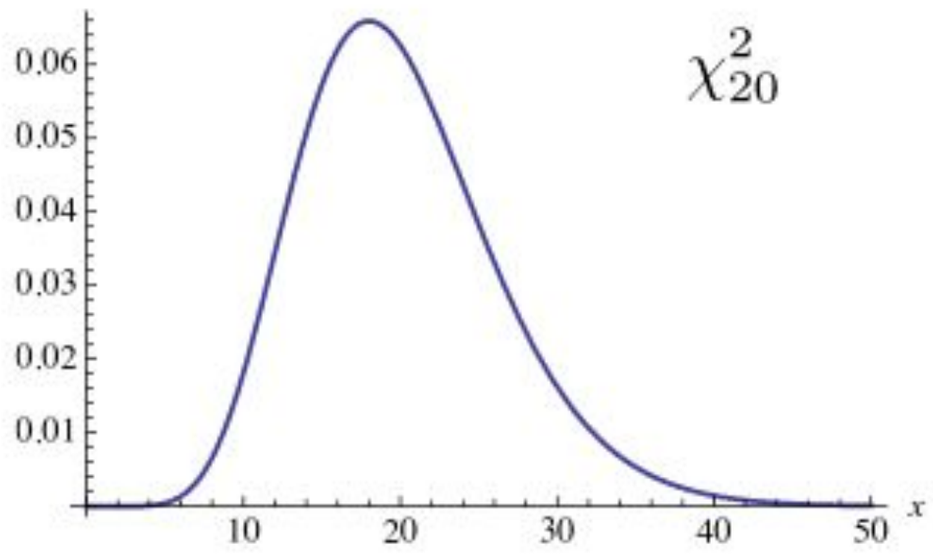
If we carry out N measurements x_i of a random variable x with variance $\sigma^2=1$ and mean $\mu=0$, then the sum

$$\sum_{i=1}^{N} x_i^2$$

follows the χ^2 (chi-square) distribution for N degrees of freedom.

χ^2 distributions



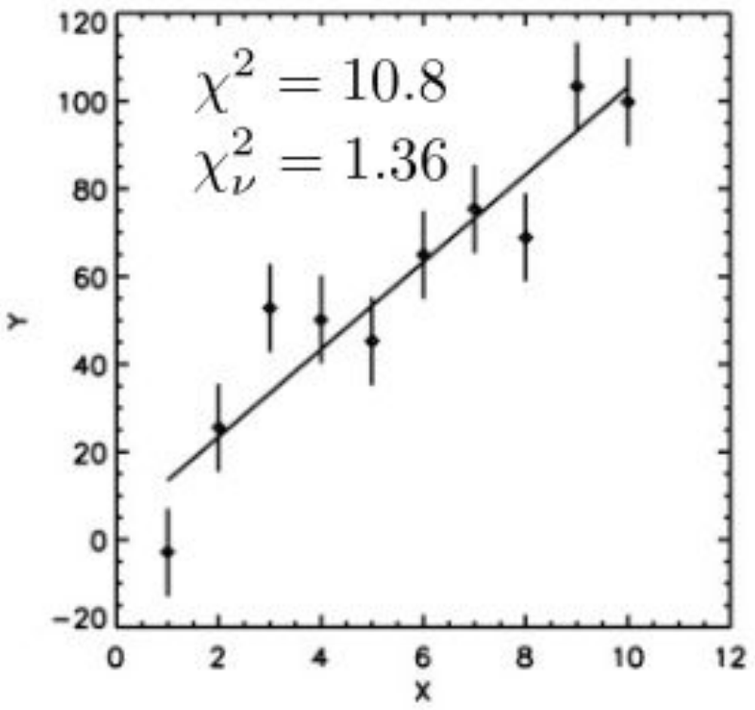


Mean of χ^2_N = N

Convenient to work with *reduced* χ^2, i.e. $\chi^2_\nu \equiv= \chi^2/N$

This always has mean 1.

Example - fitting straight line:



If each measurement y_i has error σ_{yi} and the straight line y=ax+b is the correct fit, then

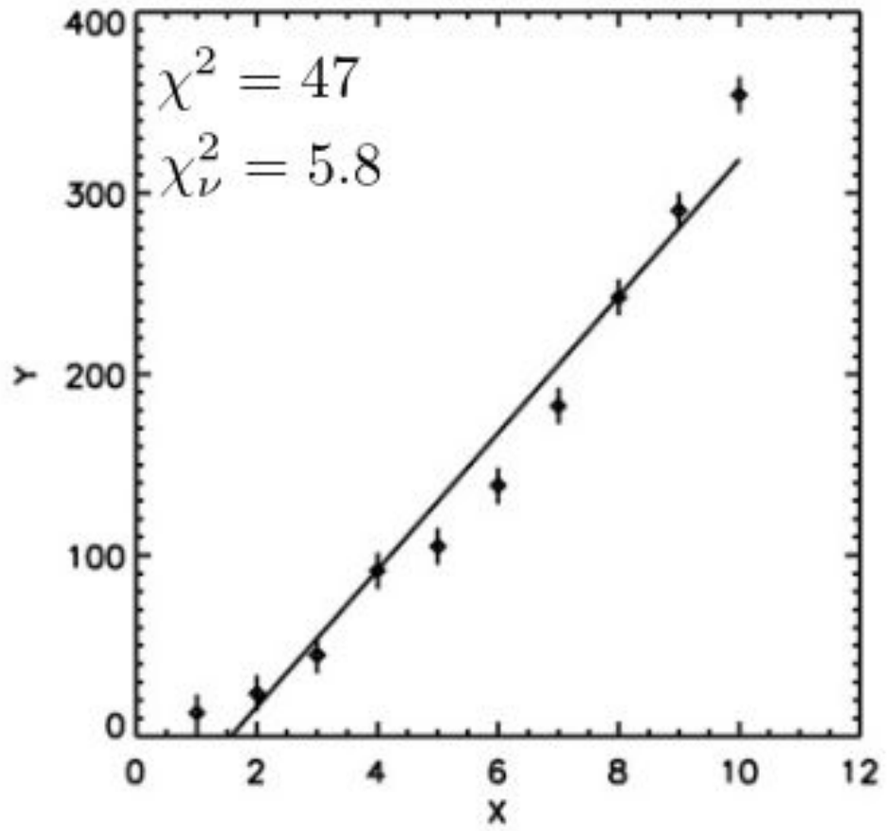
$$Y_i = \frac{y_i - (ax_i + b)}{\sigma_{y_i}}$$

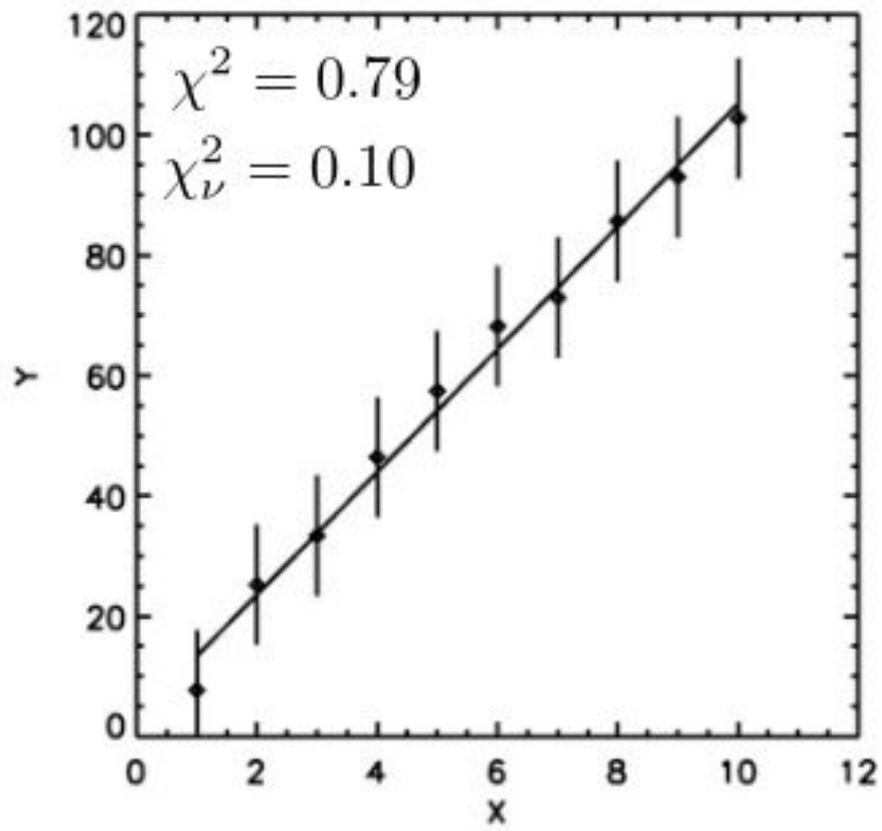
should be normally distributed with μ=0 and σ^2=1.

Hence,

$$\chi^2 = \sum_{i=1}^{N} \frac{y_i - (ax_i + b)}{\sigma_{y_i}}$$

follows the χ^2 distribution for N-2 degrees of freedom (N data, 2 free parameters).





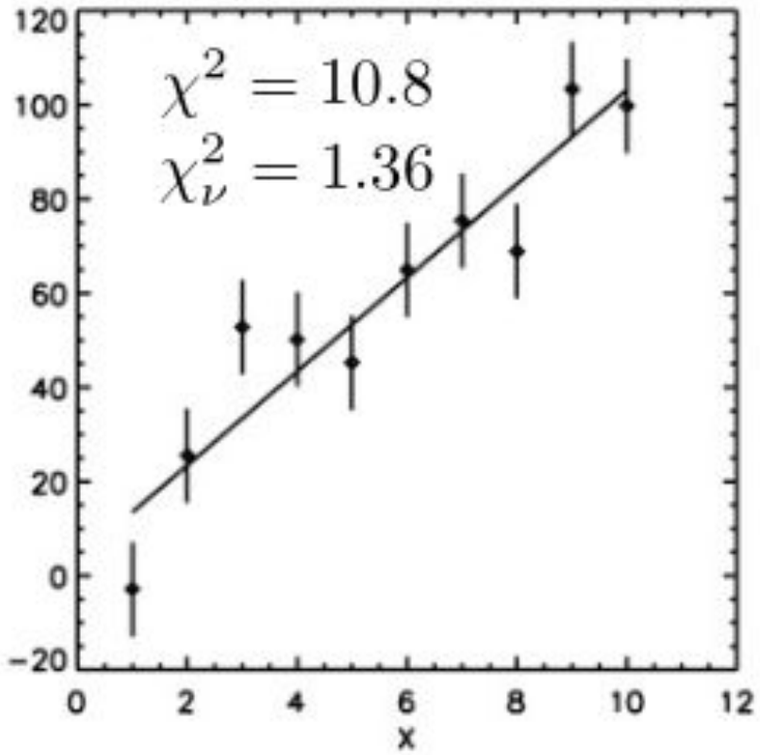
$\chi^2_\nu \gg 1$

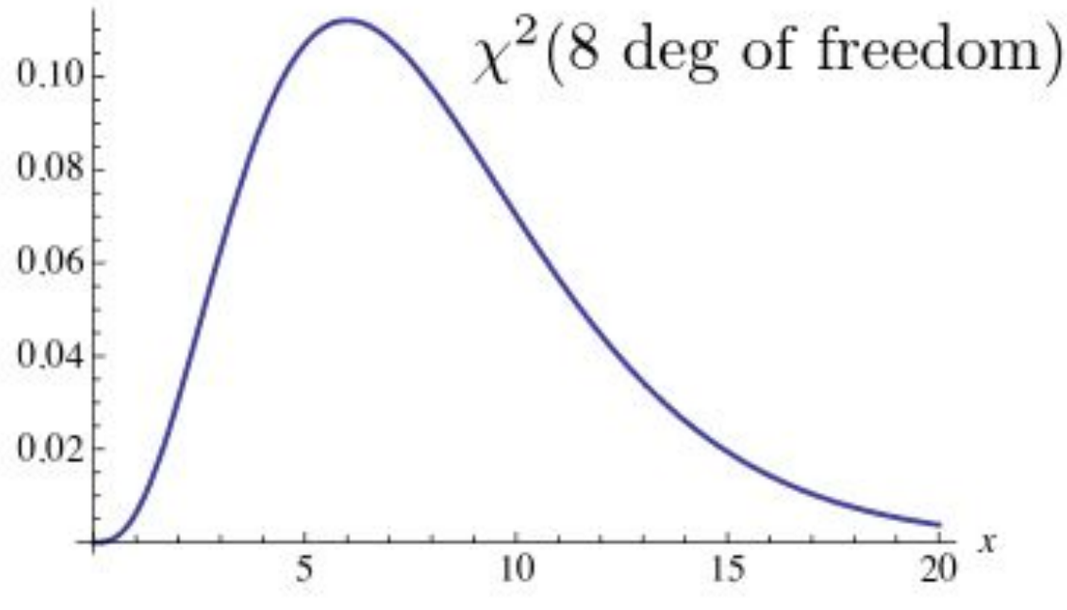
Poor fit - inappropriate model?

$\chi^2_\nu \ll 1$

Scatter too small - errors overestimated?

What is a "reasonable" range of values for χ^2_ν?



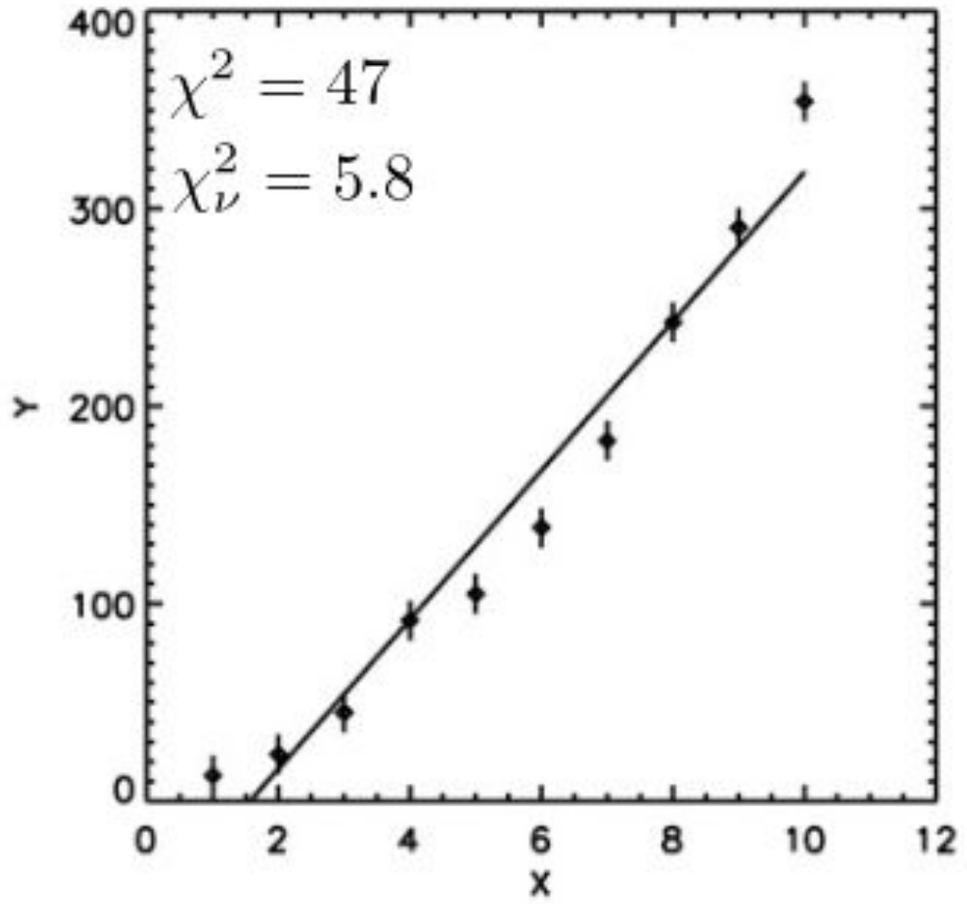


$$P(\chi^2 \geq 10.8) = \int_{10.8}^{\infty} \chi^2_8 d\chi^2 = 0.21$$

There is a 21% probability that the observed χ^2 value could have arisen out of chance.

In other words, we expect $\chi^2 > 10.8$ one out of five times the experiment is repeated.

Data are *consistent* with model assumptions (i.e. linear relation between *x* and *y*, in this case).



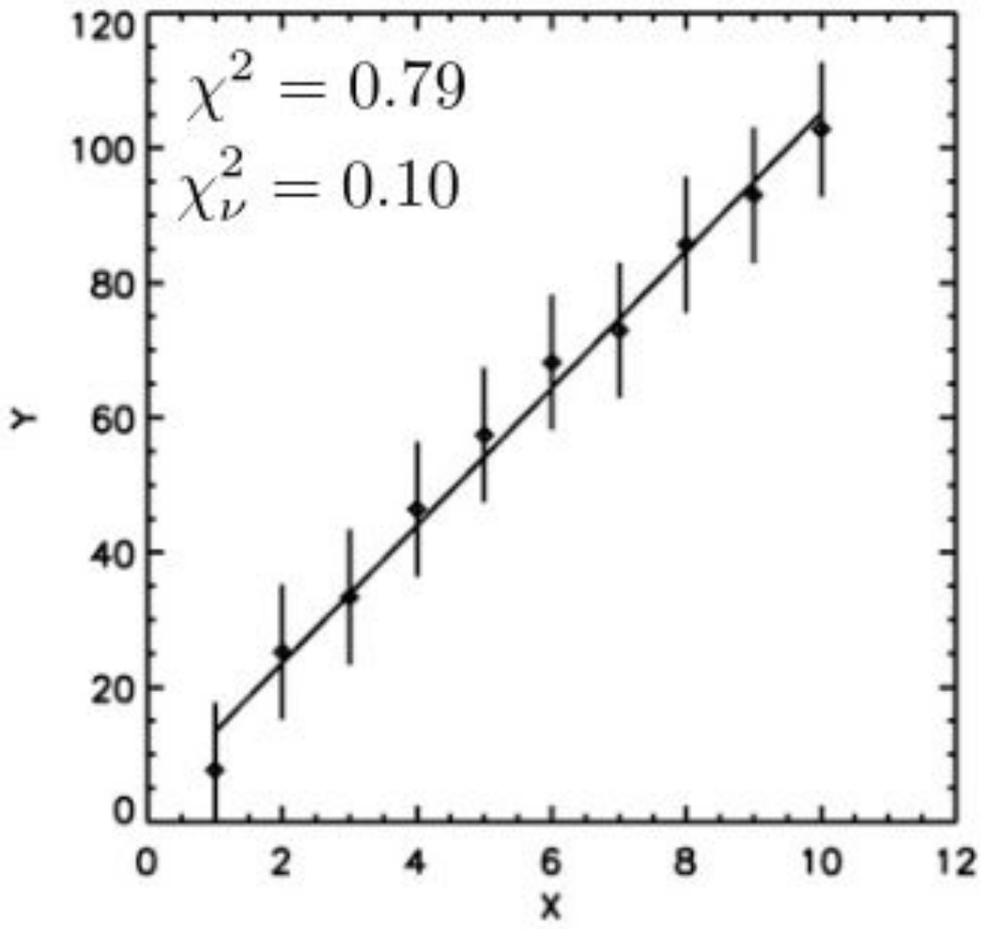
$$P(\chi^2 \geq 47) = 1.53 \times 10^{-7}$$

for 8 deg. of freedom

Frequently, one simply refers to the 'P'-value for the χ^2 (or some other) statistical test.

Small 'P' value → unlikely that the observed departures from the model could have arisen by chance. Data are *inconsistent* with the model.

A large 'P' value (~1) does not "prove" that a model is correct! It only means that the *existing* data cannot prove it wrong!



$$P(\chi^2 \leq 0.79) = 7.4 \times 10^{-4}$$

Also very unlikely that data points would line up so perfectly, considering uncertainties

Would only happen in ~1 out of 1350 cases.

The Kolmogorov-Smirnov (K-S) test

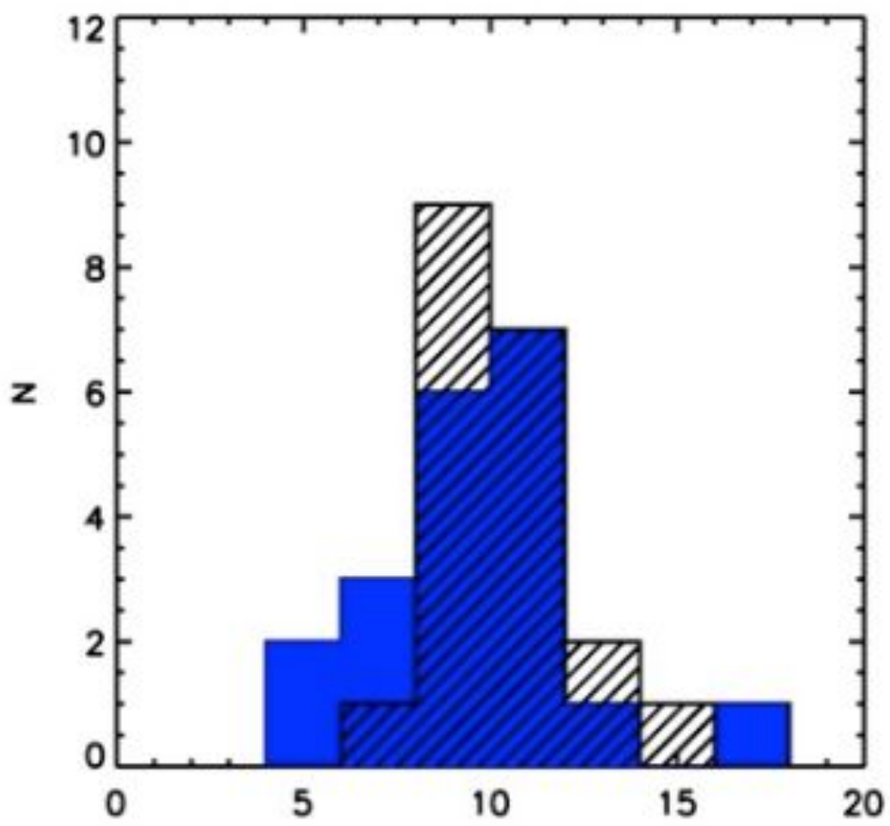
- General test to determine if two datasets "differ" (i.e. are consistent with being drawn from the same parent distribution or not).
- Can also be used to determine whether a single dataset "differs" from an assumed parent distribution
- Very general, makes no assumptions about the shape of the parent distribution

K-S test

Assume we have two data sets:

A = {9.5, 11.4, 16.1, 8.4, 9.4, 7.7, 10.6, 5.7, 7.7, 10.3, 12.0, 5.2, 8.4, 11.9, 7.3, 12.2, 8.6, 10.4, 10.6, 8.5}

B = {10.2, 9.5, 10.2, 8.6, 8.2, 10.3, 13.8, 12.8, 11.2, 10.0, 6.9, 10.2, 9.0, 9.5, 10.9, 9.6, 9.5, 11.1, 14.8, 9.6}

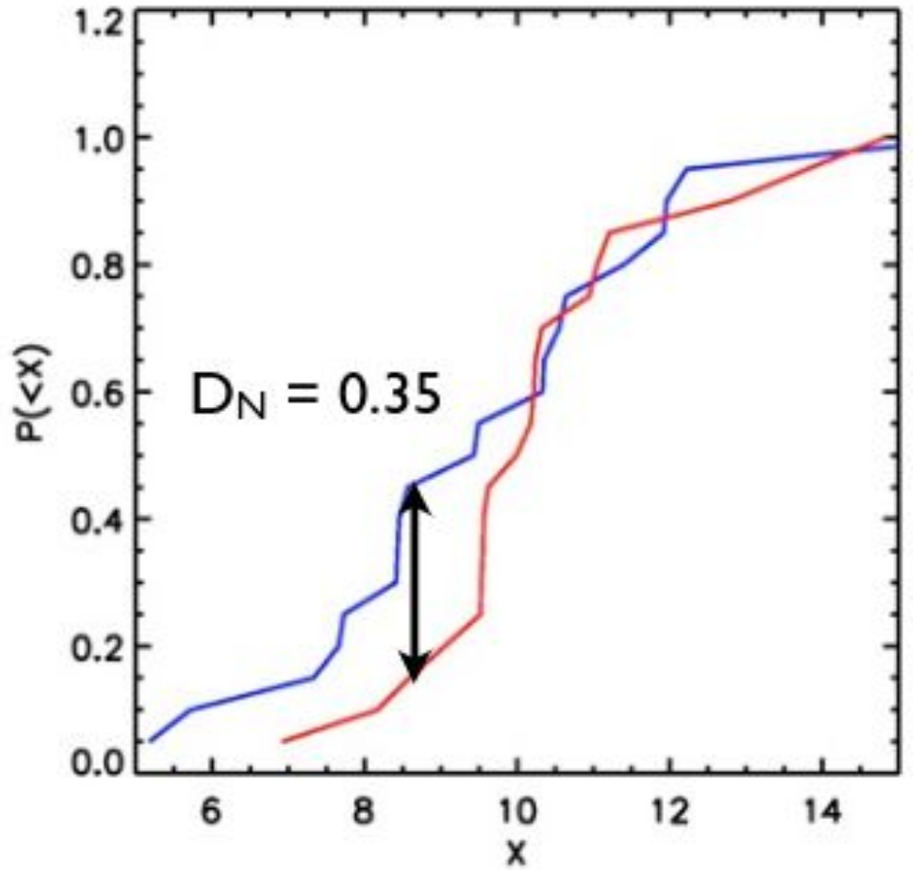


Do A and B differ?

Compare cumulative distributions of sample A and B:

Define D_N as the largest difference between the two cumulative distributions.

K-S test P value is the probability that D_N could have arisen by chance.



$$P_{KS} = 2\sum_{i=1}^{\infty}(-1)^{i+1}\exp\left(-\frac{2i^2 D_N^2}{1/n_A + 1/n_B}\right)$$

For D_N=0.35, n_A=n_B=20 we get P = 0.172, i.e. the observed D_N occurs by chance one out of 6 times.

Bayesian Inference

LIKELIHOOD
The probability of "B" being True, given "A" is True

PRIOR
The probability "A" being True. This is the knowledge.

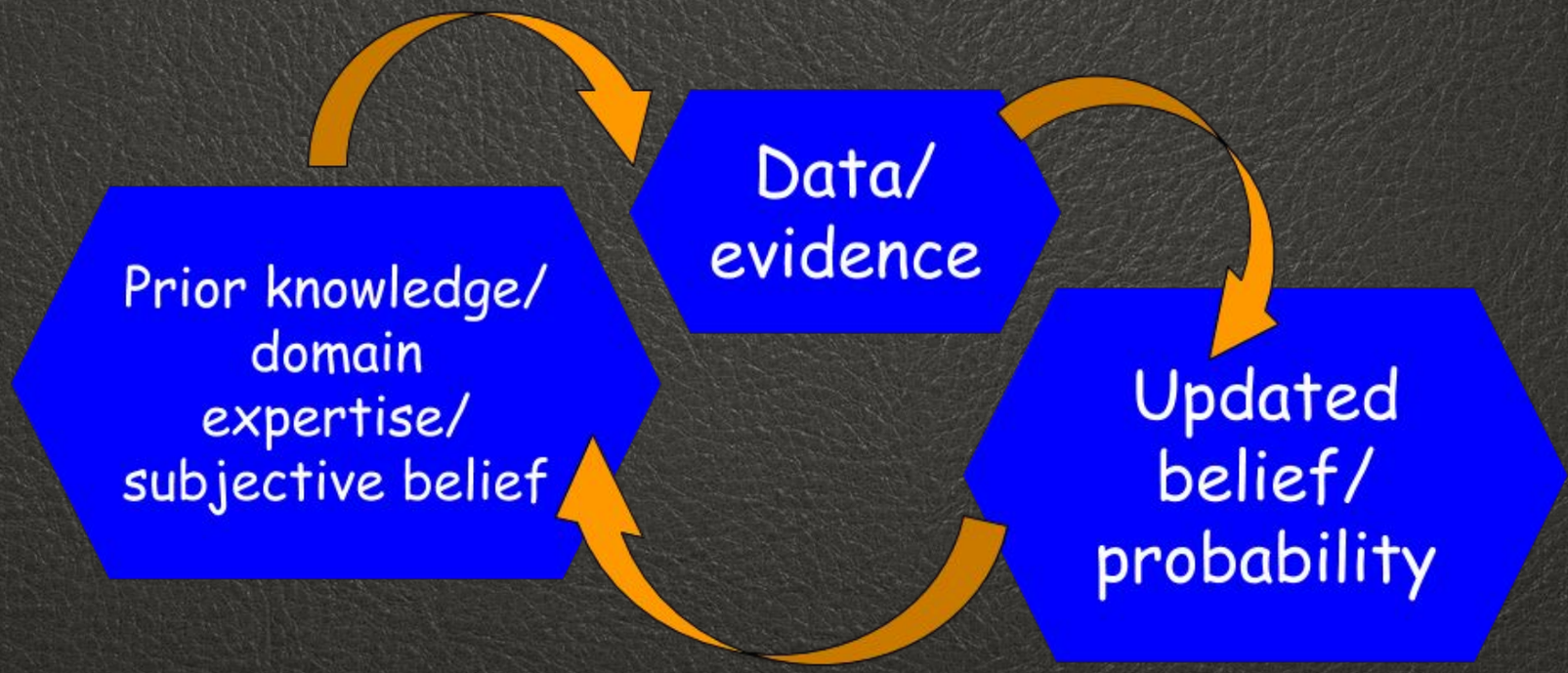
$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

POSTERIOR
The probability of "A" being True, given "B" is True

MARGINALIZATION
The probability "B" being True.

- For example, if a disease is related to age, then, using Bayes' theorem, a person's age can be used to more accurately assess the probability that they have the disease, compared to the assessment of the probability of disease made without knowledge of the person's age.

- Bayes' Theorem brings in the concept of 'subjectivity' or 'the degree of belief' into hard statistical modelings. Bayes' rule is the only mechanism that can be used to gradually update the probability of an event as the evidence or data is gathered sequentially.



Bayes Theorem Example

- Given a disease, D, with a *prevalence* 10%,
- And a test, T, with a *sensitivity* 95% and specificity 85%.

- What is the probability that subjects who test positive for D really have D?

- p(D=1) = 0.1 "prevalence"
- p(T=1 | D=1) = 0.95 "sensitivity"
- p(T=0 | D=0) = 0.85 "specificity"

$$
\begin{aligned}
p(D=1|T=1) &= \frac{p(T=1|D=1)p(D=1)}{p(T=1|D=1)p(D=1)+p(T=1|D=0)p(D=0)} \\
&= \frac{0.95 \times 0.1}{0.95 \times 0.1 + 0.15 \times 0.9} \\
&= 0.413
\end{aligned}
$$

Information Criteria

- This could be a very large topic strictly related to hypothesis testing.
- In a Bayesian scenario everything we can know given the data and the prior information is stated by the Bayes' Theorem:

$$p(M, \boldsymbol{\theta}|D, I) = \frac{p(D|M, \boldsymbol{\theta}, I)\, p(M, \boldsymbol{\theta}|I)}{p(D|I)}.$$

- And which is to be favored between two (ore more...) models M_1 and M_2 is computed by the "odd ratio":

$$O_{21} \equiv \frac{p(M_2|D, I)}{p(M_1|D, I)}.$$

Information Criteria

- The posterior probability for model M can be computed by the posterior distribution of the parameters after having marginalized (i.e. integrated) over the parameter space.

- This is a very sophisticated approach but, often, computationally hard to apply. There are simpler tools, although with limitations and assumptions.

- The BIC (Bayesian Information Criterion) assumes Gaussian posterior PDF, and becomes:

$$\mathrm{BIC} \equiv -2 \ln \left[L^0(M) \right] + k \ln N,$$

  - $L^0(M)$ is the maximum of the likelihood, k is the number of model parameters and N is the number of data point.

Information Criteria

- Another frequently applied criterion is the Akaike Information Criterion (AIC):

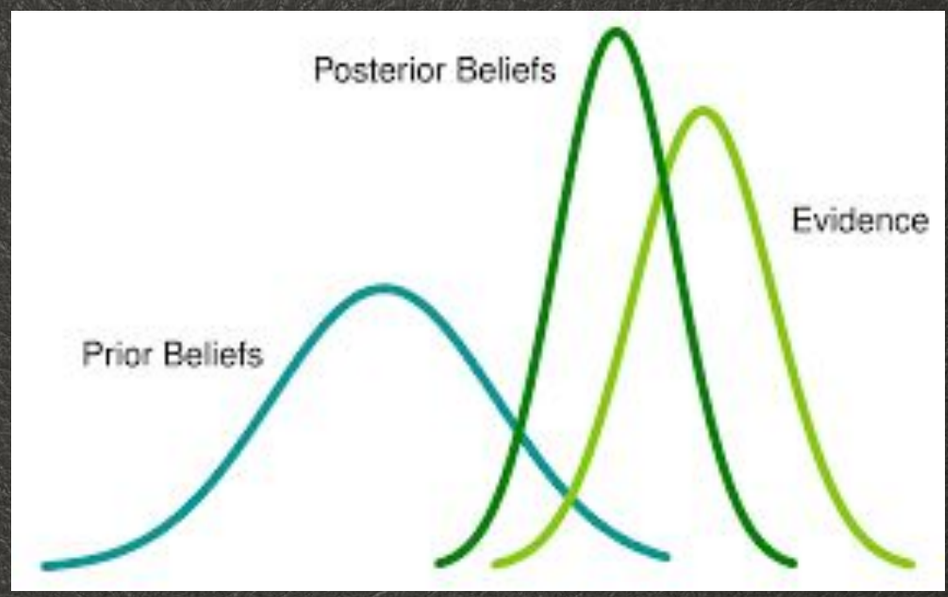
$$\mathrm{AIC} \equiv -2 \ln \left(L^0(M) \right) + 2k + \frac{2k(k+1)}{N-k-1},$$

- The AIC is a simple approach based on an asymptotic approximation.

- When multiple models are compared, the one with the smallest AIC or BIC is the best model to select. If the models are equally successful in describing the data (they have the same value of $L^0(M)$, then the model with fewer free parameters wins.

Jupyter notebook exercises:

1. FundamentalStatistics
2. Bayes



Posterior PDF Analysis

- As soon as the number of parameters grows, direct exploration of the posterior PDF becomes impractical.
- One needs sampling tools. The most widely used is the Markov Chain Monte Carlo (MCMC), yet there are several alternatives.
- Let's then assume to have our posterior PDF as:

$$p(\theta) \equiv p(M(\boldsymbol{\theta})|D, I) \propto p(D|M(\boldsymbol{\theta}), I)\, p(\boldsymbol{\theta}|I).$$

- In general we might want to estimate an integral as:

$$I(\theta) = \int g(\theta) p(\theta)\, d\theta.$$

- e.g. for marginalization $g(\theta)=1$, average $g(\theta)=\theta$, credible intervals, etc.

Markov Chain Monte Carlo

- When a direct evaluation of an integral, numerically or analytically, is impossible a Monte Carlo approach is often feasible.
- Let's generate M values of the parameter set uniformly sampled within the integration volume V_θ. The integral turns out to be:

$$I \approx \frac{V_\theta}{M} \sum_{j=1}^{M} g(\theta_j)\, p(\theta_j).$$

- The algorithm works, but it is very inefficient in particular for high-dimensional integral.

- MCMC methods return a sample of points, or chain, from the *k*-dimensional parameter space, with a distribution that is asymptotically proportional to $p(\theta)$.

Markov Chain Monte Carlo

- With such a chain the integral becomes:

$$I = \frac{1}{M} \sum_{j=1}^{M} g(\theta_j).$$

- E.g., to estimate the expectation value for θ (i.e., $g(\theta) = \theta$), we simply take the mean value of all θ in the chain.
- A Markov chain is a sequence of random variables where a given value depends *only on its preceding value*. Given the *present* value, past and future values are independent.
- The process generating such a chain is called the Markov process.

Markov Chain Monte Carlo

- The process generating such a chain is called the Markov process and can be described as:

$$p(\theta_{i+1}|\{\theta_i\}) = p(\theta_{i+1}|\theta_i),$$

- To reach an equilibrium, or stationary, distribution of positions, it is necessary that the transition probability be symmetric:

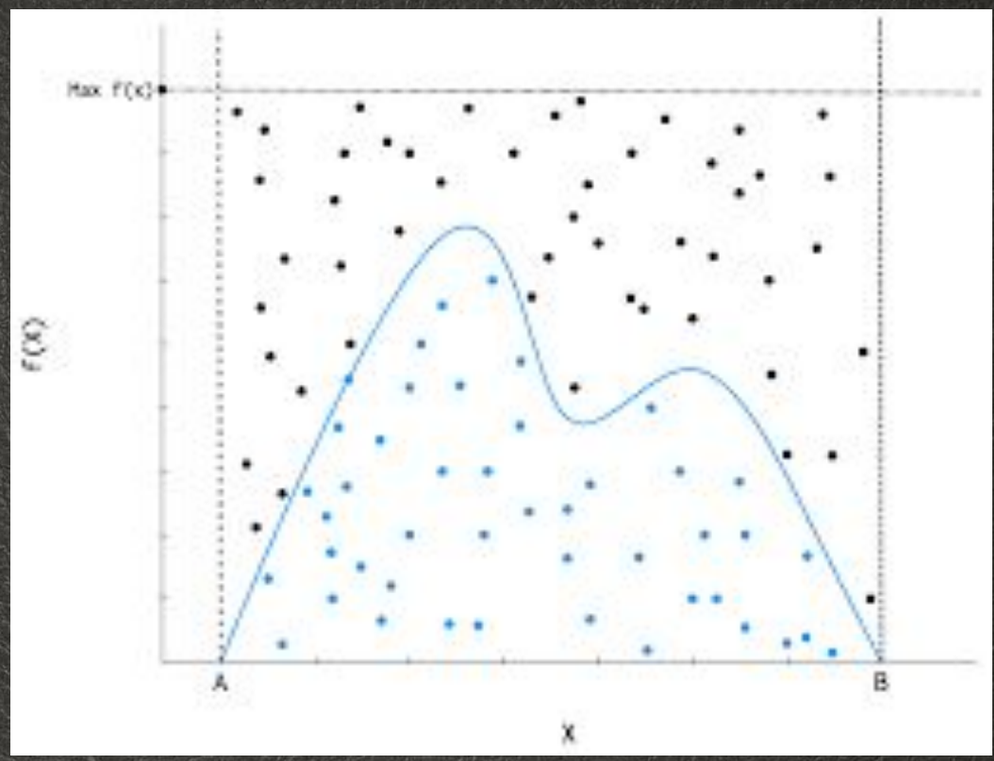
$$p(\theta_{i+1}|\theta_i) = p(\theta_i|\theta_{i+1}).$$

- There are various algorithms for producing Markov chains that reach some prescribed equilibrium distribution, $p(\theta)$.

- Interactive demo: https://chi-feng.github.io/mcmc-demo/

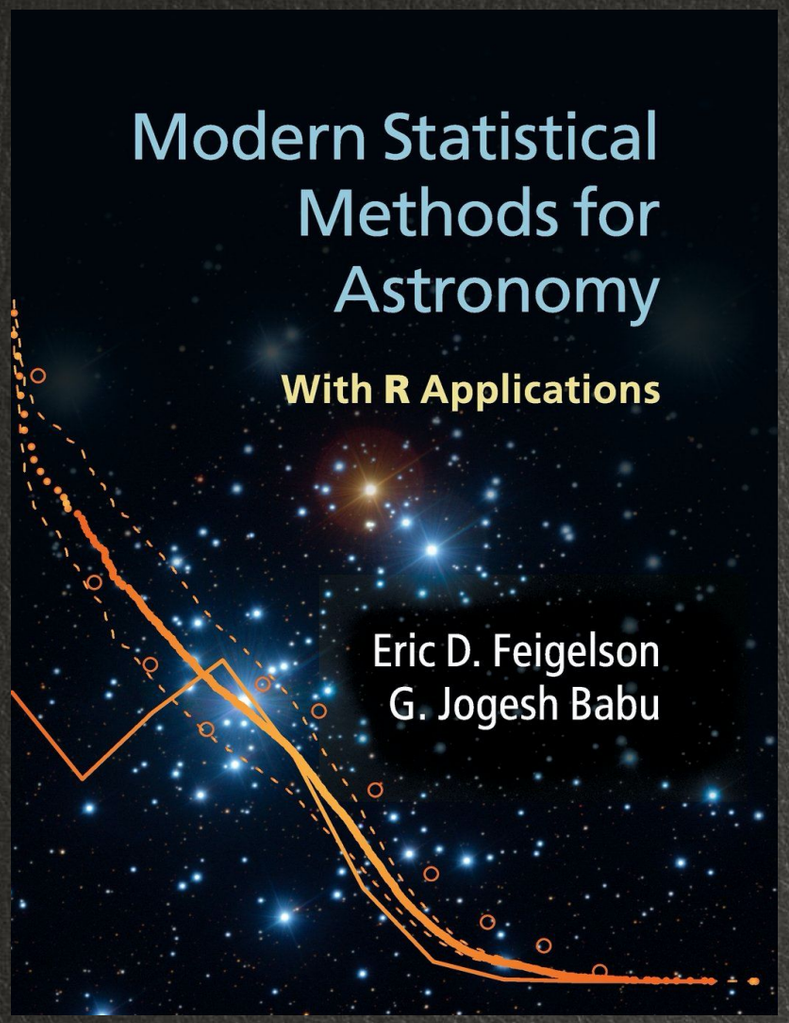
Jupyter notebook (optional) homework:

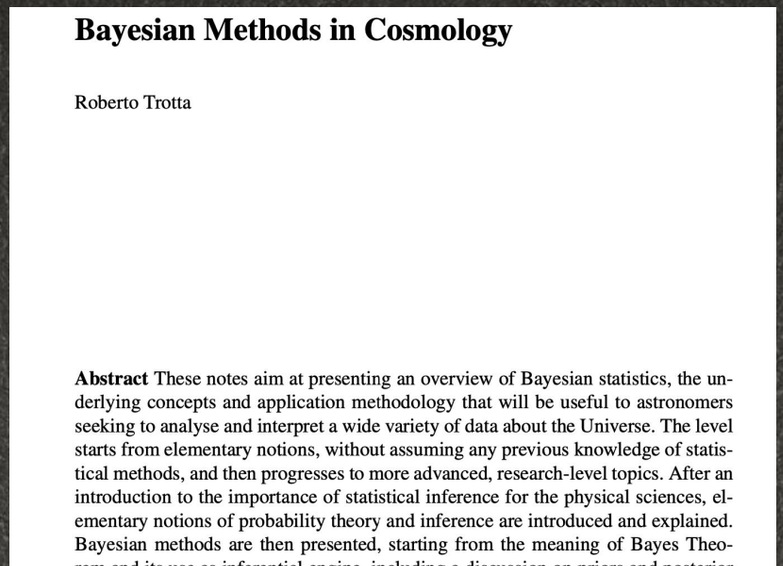
1. FreqBayes
2. FreqBayes2
3. FreqBayes3
4. FreqBayes4
5. FreqBayes5



- Please follow them carefully. They are going to give a feeling about Bayesian inference better than many textbook!

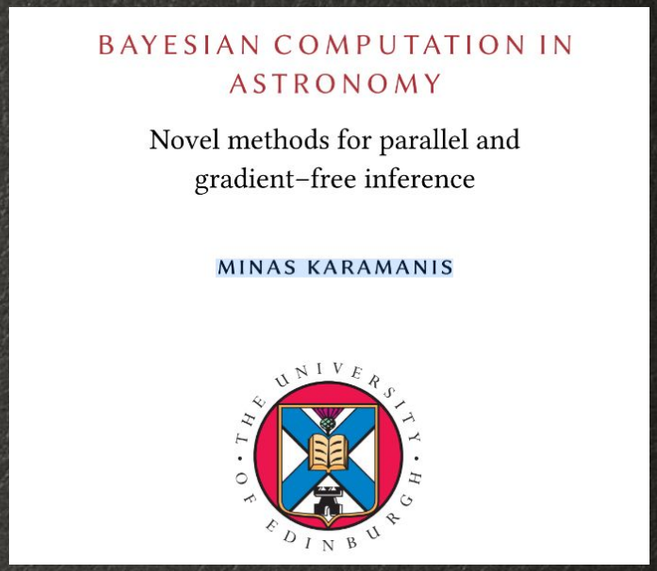
REFERENCES AND DEEPENING




Bayesian Methods in Cosmology

Roberto Trotta

Abstract These notes aim at presenting an overview of Bayesian statistics, the underlying concepts and application methodology that will be useful to astronomers seeking to analyse and interpret a wide variety of data about the Universe. The level starts from elementary notions, without assuming any previous knowledge of statistical methods, and then progresses to more advanced, research-level topics. After an introduction to the importance of statistical inference for the physical sciences, elementary notions of probability theory and inference are introduced and explained. Bayesian methods are then presented, starting from the meaning of Bayes Theorem and its use as inferential engine, including a discussion on priors and posterior distributions. Numerical methods for generating samples from arbitrary posteriors (including Markov Chain Monte Carlo and Nested Sampling) are then covered. The last section deals with the topic of Bayesian model selection and how it is used to assess the performance of models, and contrasts it with the classical p-value approach. A series of exercises of various levels of difficulty are designed to further the understanding of the theoretical material, including fully worked out solutions for most of them.


BAYESIAN COMPUTATION IN ASTRONOMY

Novel methods for parallel and gradient–free inference

MINAS KARAMANIS

THE UNIVERSITY OF EDINBURGH

Roberto Trotta

Minas Karamanis

Eric Feigelson & Jogesh Babu