Lecture 1 :
Basic descriptive statistics

Lecture 1 : basic descriptive statistics

- What is the point of statistics?
- Common pitfalls
- Mean, variance, median of a sample, and their errors
- Probability distributions (binomial, Poisson, Gaussian)
- Error propagation (linear, non-linear)
- Optimal combination of data

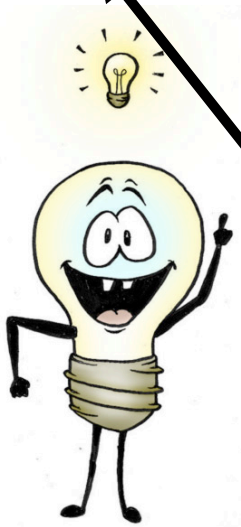
The process of science

Design a question

Obtain measurements

Analyze data

Conclude (test hypothesis)



The point of statistics

"If your experiment needs statistics, you ought to have done a better experiment" [E.Rutherford]

"A body of methods for making wise decisions in the face of uncertainty" [W.Wallis]

- It allows us to formulate the logic of what we are doing and why. It allows us to make precise statements.
- It allows us to quantify the uncertainty in any measurement, which should always be stated.
- It allows us to avoid pitfalls such as confirmation bias (distortion of conclusions by preconceived beliefs)

Common uses of statistics

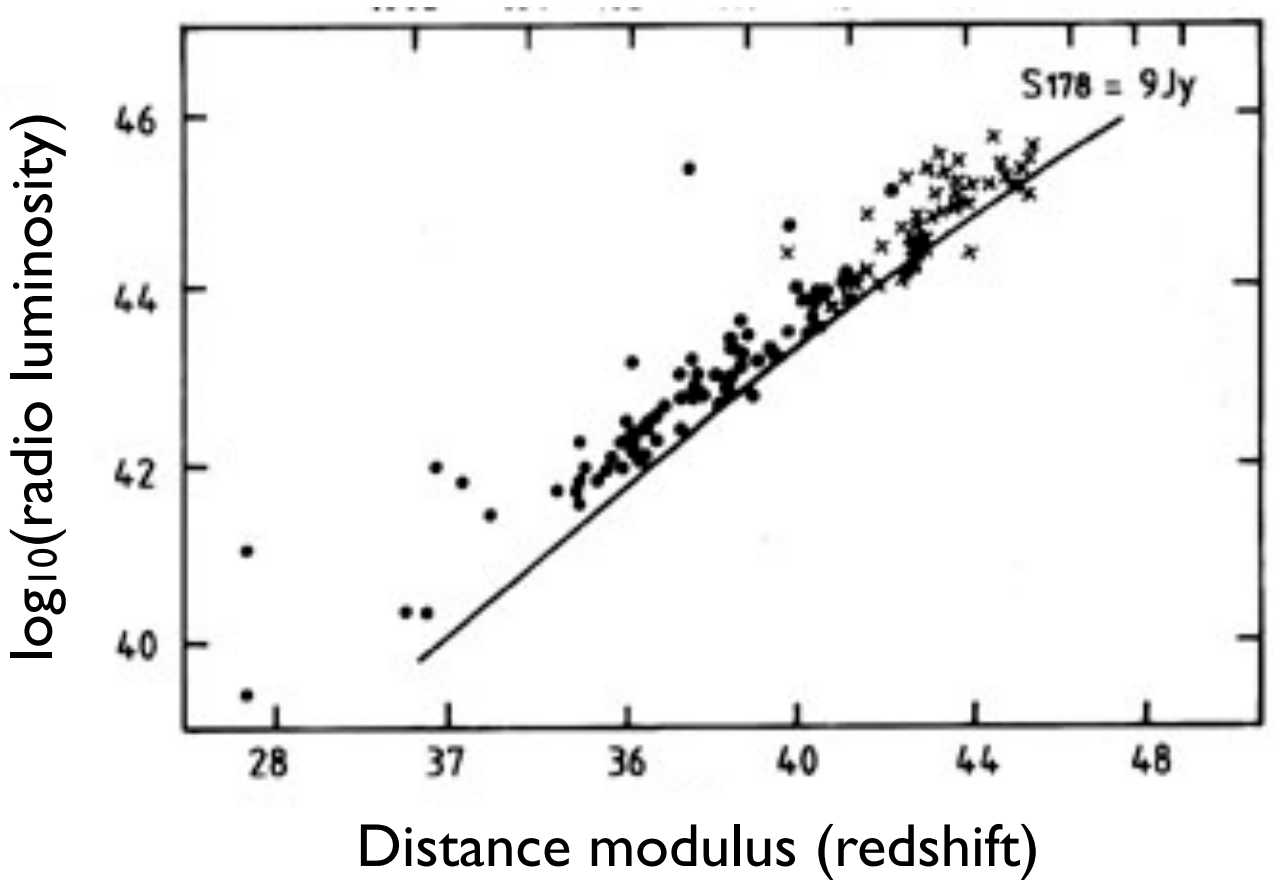
- Measuring a quantity ("parameter estimation") : given some data, what is our best estimate of a particular parameter? What is the uncertainty in our estimate?

- Searching for correlations : are two variables we have measured correlated with each other, implying a possible physical connection?

- Testing a model ("hypothesis testing") : given some data and one or more models, are our data consistent with the models? Which model best describes the data?

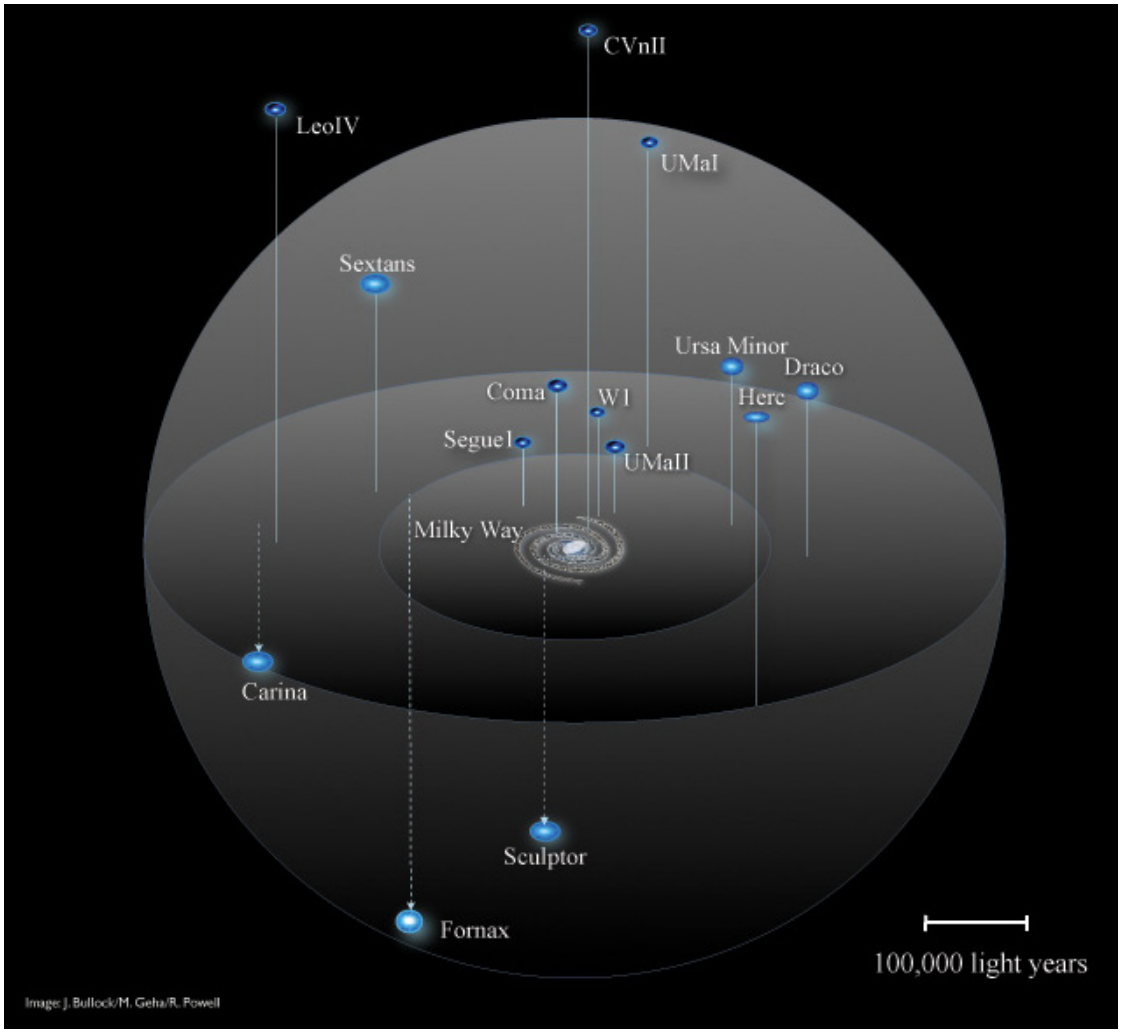
Common statistical issues for astronomers

- Selection effects leading to spurious correlations, for example Malmquist bias



Common statistical issues for astronomers

- Small samples leading to noisy results



Common statistical issues for astronomers

- Confirmation bias : conclusions distorted by our pre-conceived idea about what the result should be

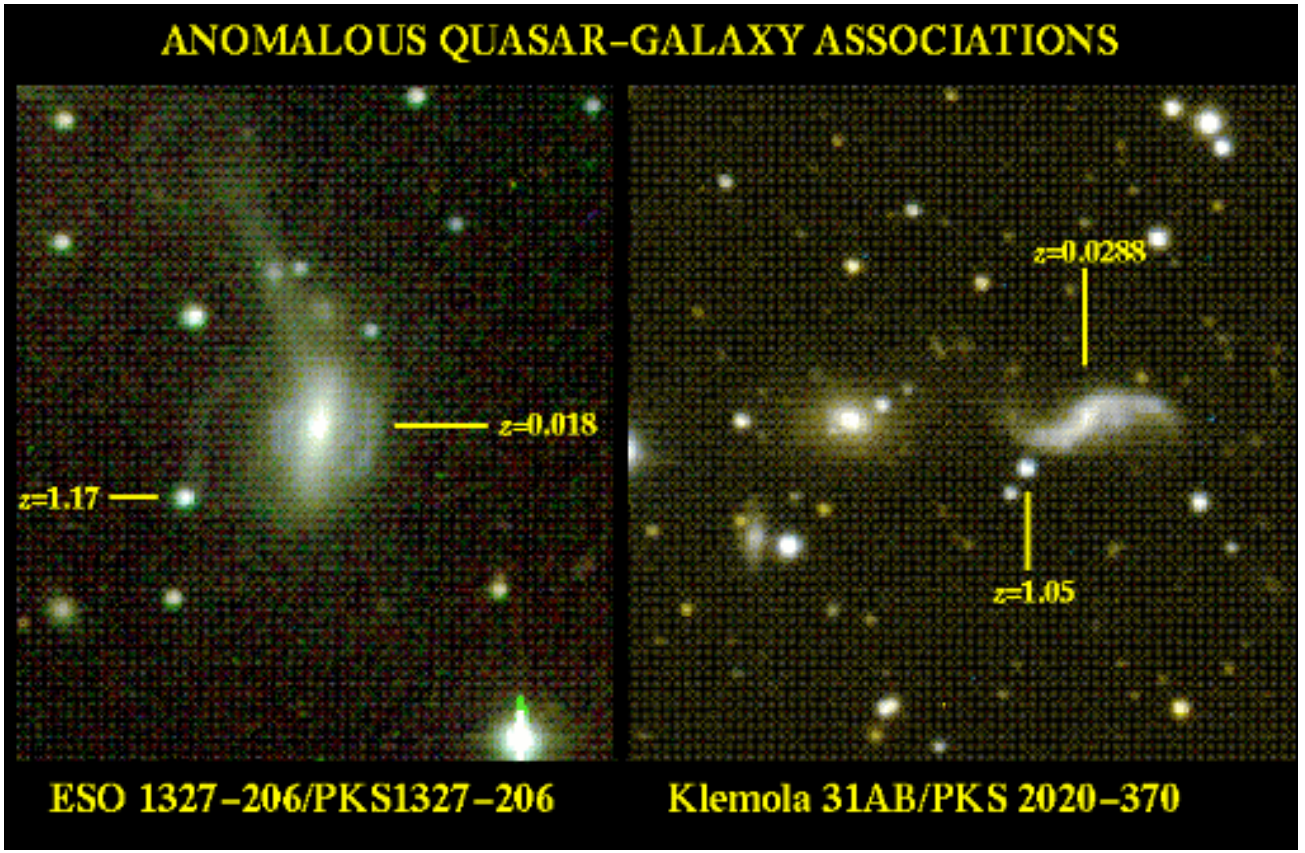
Title: On the measurement of cosmological parameters

Authors: Rupert A. C. Croft, Matthew Dailey (CMU)
(Submitted on 14 Dec 2011)

Abstract: We have catalogued and analysed cosmological parameter determinations and their error bars published between the years 1990 and 2010. Our study focuses on the number of measurements, their precision and their accuracy. The accuracy of past measurements is gauged by comparison with the WMAP7 results. The 637 measurements in our study are of 12 different parameters and we place the techniques used to carry them out into 12 different categories. We find that the number of published measurements per year in all 12 cases except for the dark energy equation of state parameter w_0 peaked between 1995 and 2004. Of the individual techniques, only BAO measurements were still rising in popularity at the end of the studied time period. The fractional error associated with most measurements has been declining relatively slowly, with several parameters, such as the amplitude of mass fluctutations sigma_{8} and the Hubble constant H_0 remaining close to the 10% precision level for a 10-15 year period. The accuracy of recent parameter measurements is generally what would be expected given the quoted error bars, although before the year 2000, the accuracy was significantly worse, consistent with an average underestimate of the error bars by a factor of ~2. When used as complement to traditional forecasting techniques, our results suggest that future measurements of parameters such as fNL, and w_a will have been informed by the gradual improvment in understanding and treatment of systematic errors and are likely to be accurate. However, care must be taken to avoid the effects of confirmation bias, which may be affecting recent measurements of dark energy parameters. For example, of the 28 measurements of Omega_Lambda in our sample published since 2003, only 2 are more than 1 sigma from the WMAP results. Wider use of blind analyses in cosmology could help to avoid this.

Common statistical issues for astronomers

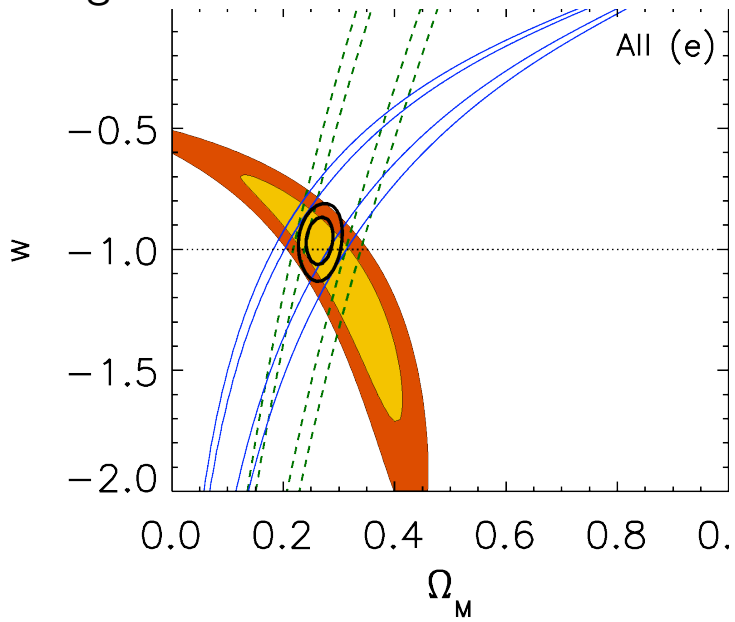
- Using the same dataset which motivated a hypothesis to test that hypothesis ("a posteriori" statistics)



Common statistical issues for astronomers

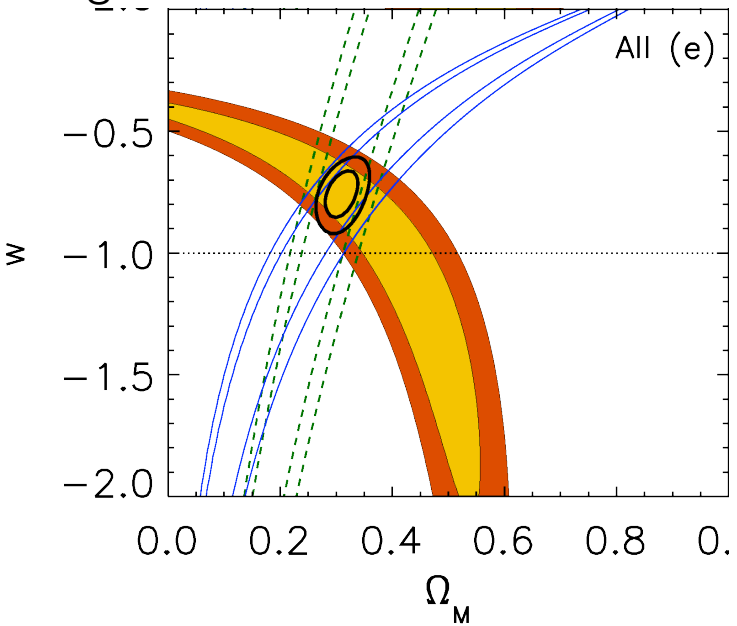
- Results dominated by systematic errors rather than by statistical uncertainties

Using **SALT2**



$w = -0.96 \pm 0.06(\mathrm{stat}) \pm 0.12(\mathrm{sys})$

Using **MLCS2k2**



$w = -0.76 \pm 0.07(\mathrm{stat}) \pm 0.11(\mathrm{sys})$

Estimating basic statistics

- Example 1 : The significance of a certain conclusion depends very strongly on whether the most luminous known quasar is included in the dataset. The object is legitimately in the dataset in terms of the pre-stated selection criteria. Is the conclusion robust?

- The conclusion may not be robust because it is dominated by a single outlier which may not be typical of the sample as a whole.

Estimating basic statistics

- A statistic is a quantity which summarizes our data
- I have a sample of N independent estimates x_i of some quantity, how can I summarize them?
- Mean (typical value) : $\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$
- Median (middle value when ranked)
- Variance (spread) closely related to standard deviation :

$$\mathrm{Var}(x) = \sigma^2(x) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2$$

Watch out for N-1 factor!

Estimating basic statistics

- We can quote an error in each of these statistics
- Error in the mean : $\sigma(\overline{x}) = \dfrac{\sigma(x)}{\sqrt{N}}$
- Error in the median : $\sigma[\mathrm{Med}(x)] = 1.25\,\dfrac{\sigma(x)}{\sqrt{N}}$
- Error in the variance : $\sigma[\mathrm{Var}(x)] = \mathrm{Var}(x)\sqrt{\dfrac{2}{N-1}}$
- Small print : the error in the mean relation holds independently of the probability distribution of x , the other two relations assume a Gaussian distribution

Estimating basic statistics

- Example 2 : we have N=10 measurements of a variable x_i = (7.6, 5.8, 8.0, 6.9, 7.2, 7.5, 6.4, 8.1, 6.3, 7.0). Estimate the mean, variance and median of this variable. What are the errors in your estimates?

- Mean = 7.08 +/- 0.24
- Variance = 0.57 +/- 0.27
- Median = 7.10 +/- 0.30

The meaning of an error bar

$$H_0 = 70 \pm 5 \ \mathrm{km\,s^{-1}\,Mpc^{-1}}$$

- What does this statement mean?
- It almost never means "H_0 is between 65 and 75"
- It almost always means "there is a 68% probability that H_0 lies in the confidence region $65 < H_0 < 75$"
- It often means "the probability distribution for H_0 is a Gaussian with mean 70 and standard deviation 5"
- So what is a probability distribution?

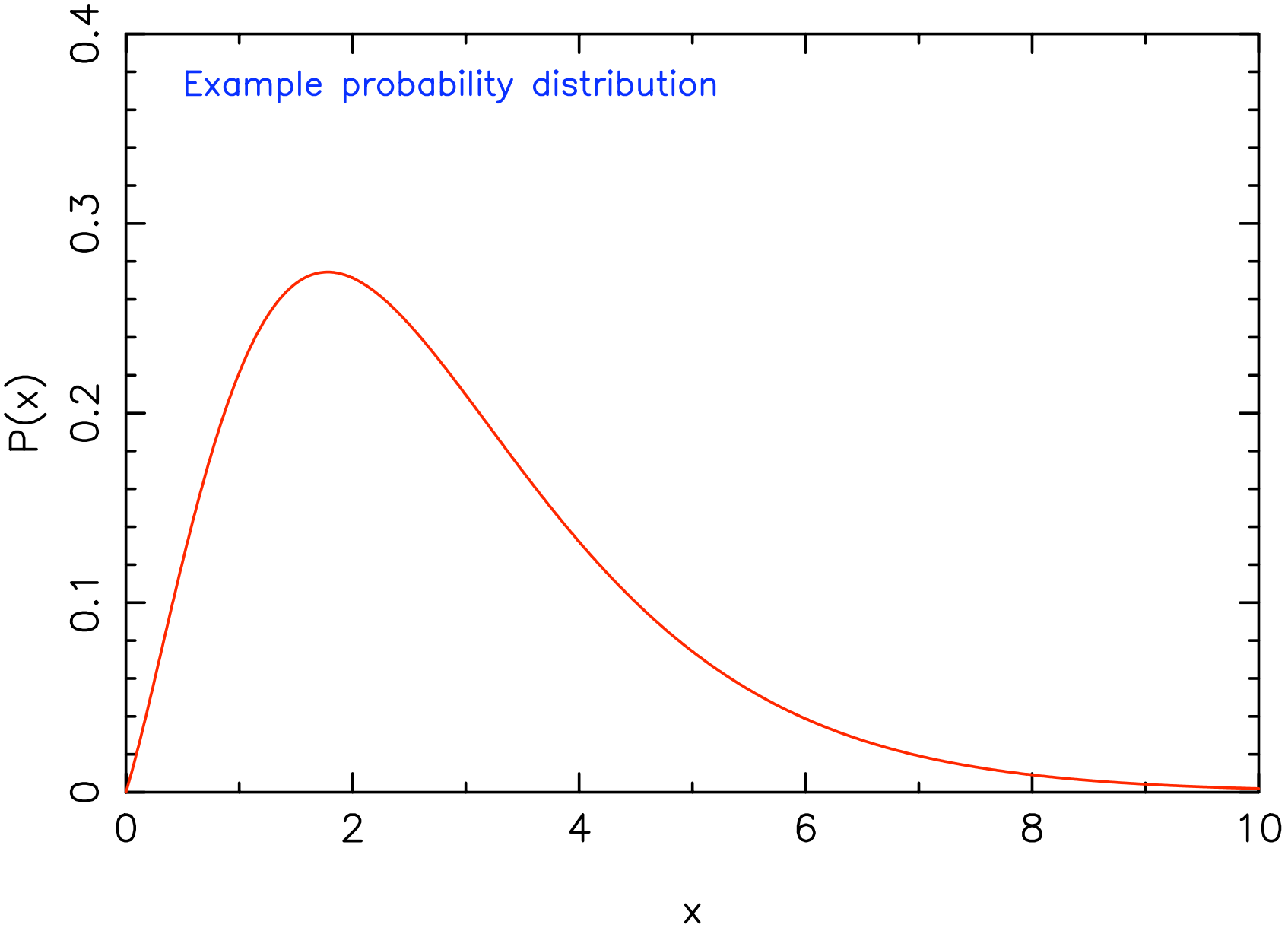
Probability distributions

- A probability distribution is a function which assigns a probability for each particular value (or range of values) of a variable x
- Must be normalized : $\int_{-\infty}^{\infty} p(x)\,dx = 1$
- Probability in range $[x_1,x_2]$ = $\int_{x_1}^{x_2} p(x)\,dx$
- Distribution may be quantified by its mean & variance:

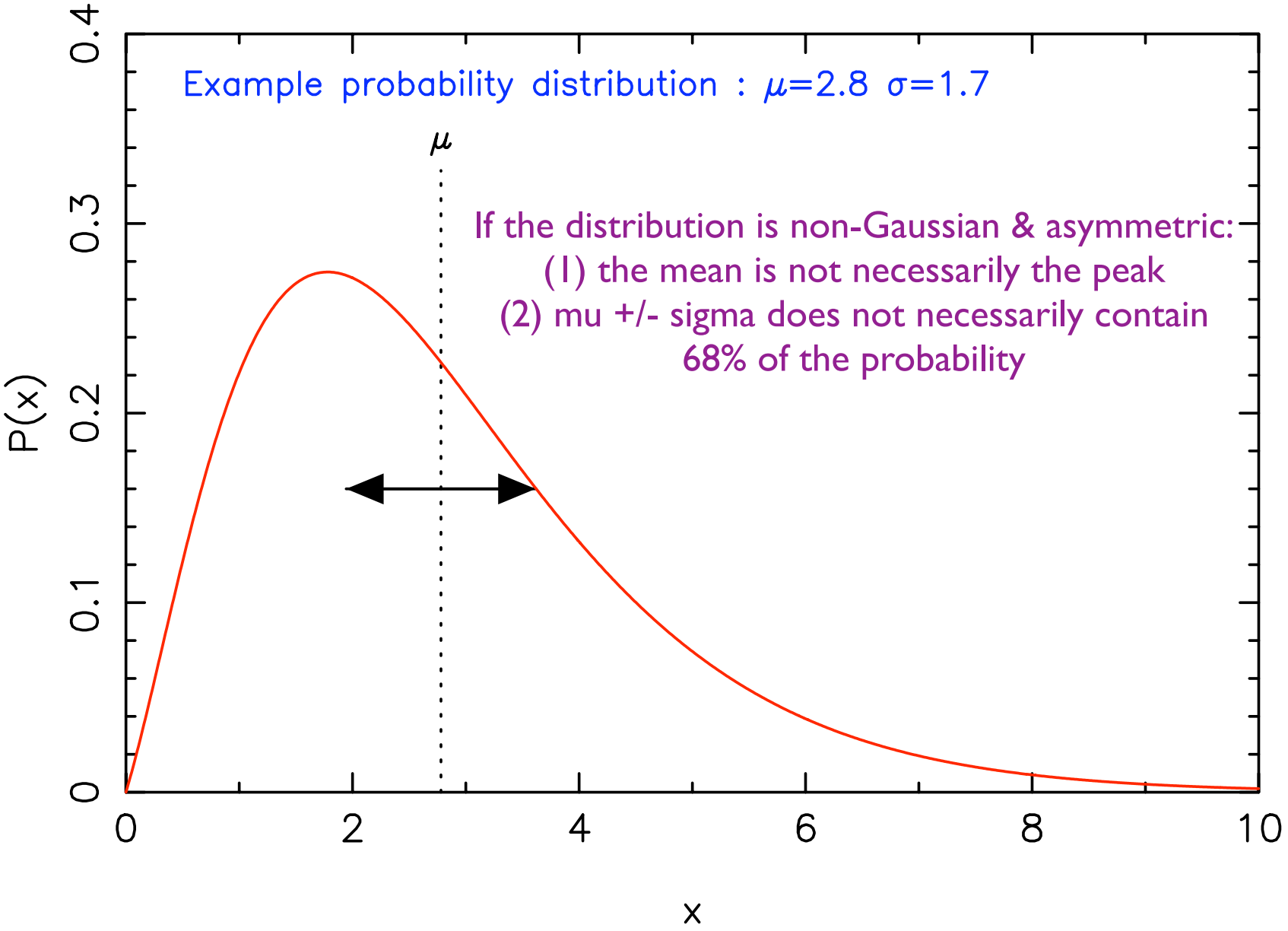
$$\mu = \overline{x} =< x >= \int_{-\infty}^{\infty} x\,p(x)\,dx$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2\,p(x)\,dx = \overline{x^2} - \overline{x}^2$$

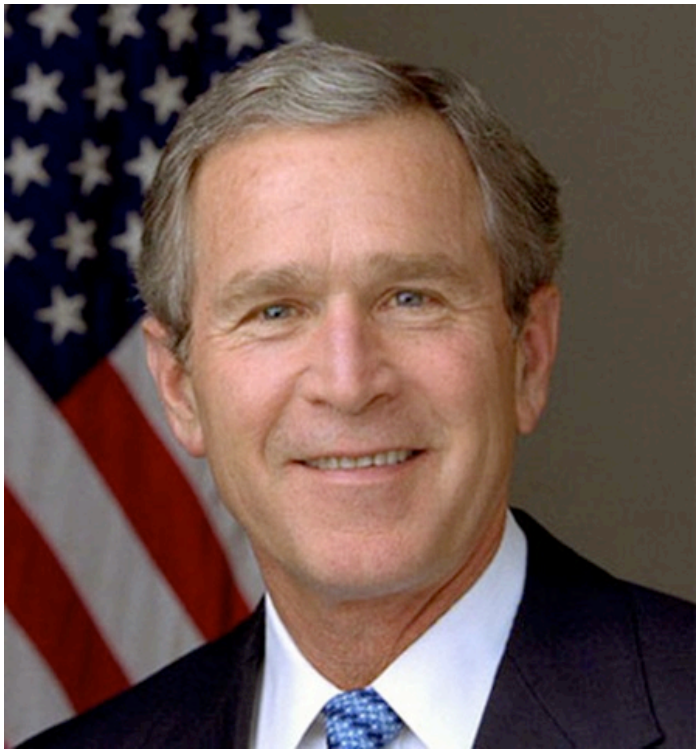
Probability distributions



Probability distributions



Lies, damn lies and statistics



Example 1

"92 million Americans will receive an average tax cut of $1083" [in 2003]

Why was this poor statistics? An average is a poor summary of the underlying probability distribution, which was heavily skewed such that the top 1% of income earners were gifted $30,127 !

Probability distributions

- Usually difficult to deduce the full probability distribution of a variable from limited observations
- Certain types of variables have well-known distributions:
- Binomial distribution
- Poisson distribution
- Gaussian or Normal distribution

Binomial distribution

- Applies in problems where there is a random process with 2 possible outcomes with probabilities p and 1-p

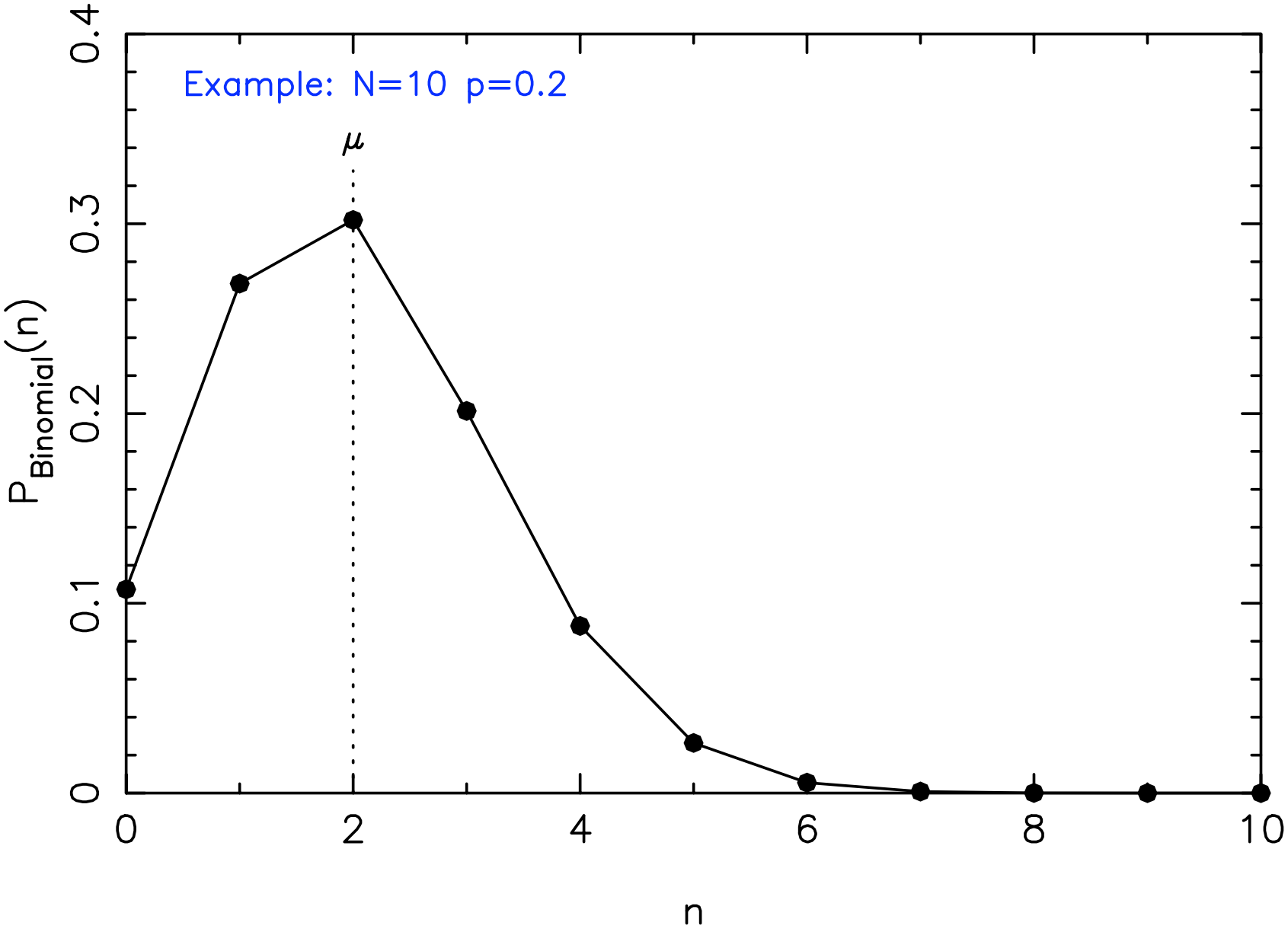
- Example : tossing a coin

$$P_{\text{binomial}}(n) = \frac{N!}{n!\,(N-n)!}\,p^n(1-p)^{N-n}$$

$N = \text{no.trials}$ $n = \text{no.successes}$ $p = \text{probability of success}$

$$\overline{n} = pN \qquad \text{Var}(n) = Np(1-p)$$

Binomial distribution



Binomial distribution

- Example 3 : I observe 100 galaxies, 30 of which are AGN. What is the best estimate of the AGN fraction and its error?

- AGN fraction = p = 30/100 = 0.3

- There are 2 possible outcomes ("AGN" or "not an AGN") so the binomial distribution applies

- Error in AGN fraction = sqrt[N p (1-p)]/N = 0.046
  [compare Poisson error = sqrt(30)/100 = 0.055]

Binomial distribution

- Example 4 : In the HST guide star catalogue, 60% of the objects are binary stars. How large a sample should be chosen to ensure that the probability of the sample containing at least 2 non-binary stars is at least 99%?

- There are 2 possible outcomes ("binary" or "non-binary") so the binomial distribution applies with p=0.6

- $P(0) + P(1) = 0.6^N + N\ 0.4\ 0.6^{(N-1)}$

- $P(0) + P(1) < 0.01$ if $N > 14$

Poisson distribution

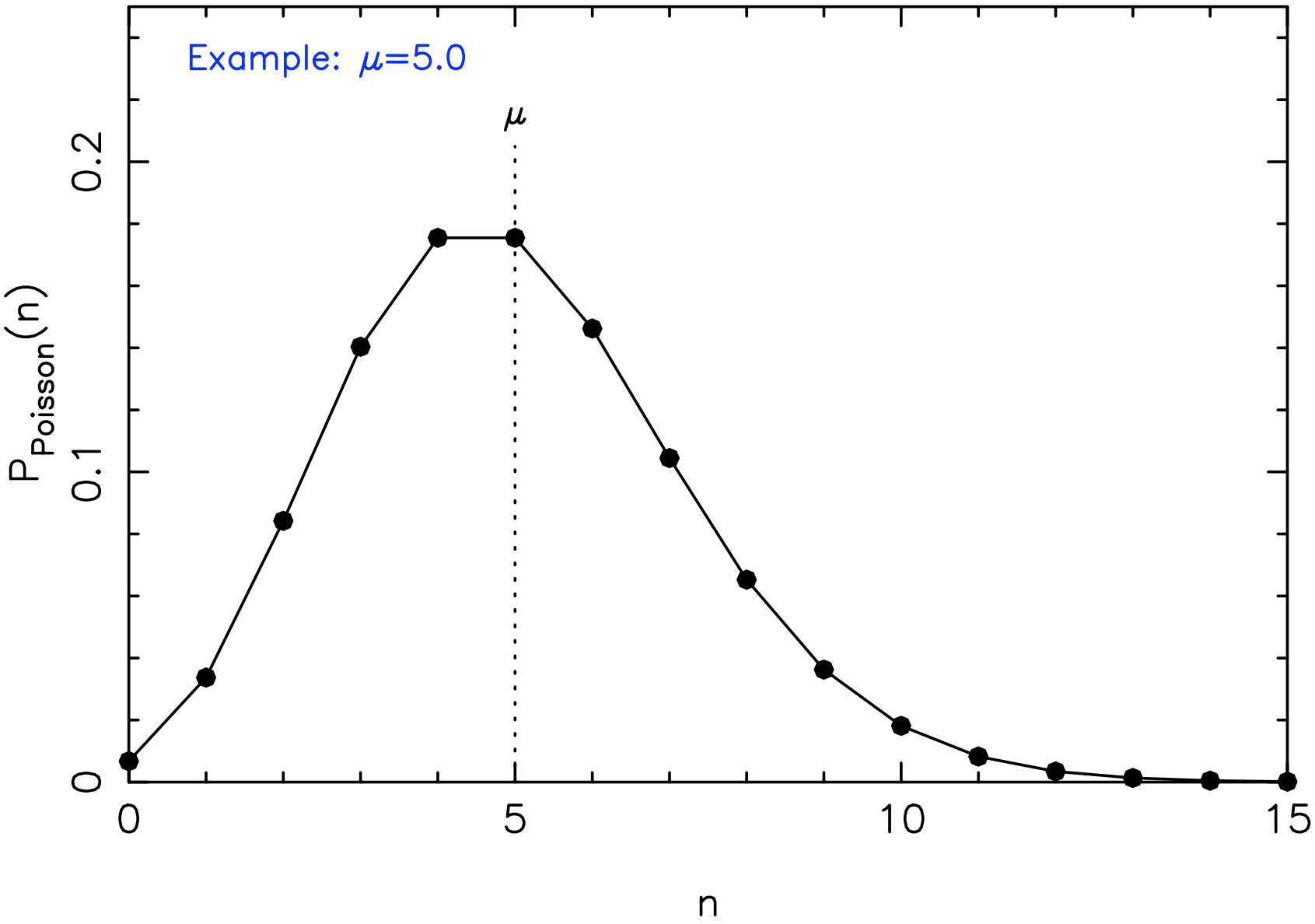
- Applies to a discrete random process where we are counting something in a fixed interval
- Example : radioactive decay , photons arriving at CCD
- The ultimate limit to any counting experiment

$$P_{\text{Poisson}}(n) = \frac{\mu^n \exp(-\mu)}{n!}$$

$n = \text{no.events observed} \quad \mu = \text{mean no.events expected}$

$$\overline{n} = \text{Var}(n) = \mu$$

Poisson distribution



The "Poisson error"

$$y = N \pm \sqrt{N}$$

- If an individual bin of data contains N events (e.g. CCD pixel contains N photons) we often place a Poisson error sqrt(N) in that bin
- Assumes the mean count is the observed count
- Bad approximation for low numbers (e.g. N=0) and in such cases the distribution is not Gaussian
- Bad approximation if the fluctuations are dominated by other processes (e.g. read noise , galaxy clustering)

Poisson distribution

- Example 5 : the density of quasars on the sky is known to be 20 per deg^2. What area of sky would we need to survey to ensure a 99% chance of finding a quasar?

- We are counting random events compared to a mean, so the Poisson distribution applies

- Mean number of quasars in area A deg^2 = 20 A

- P(0) = exp(-20 A) = 0.01 if A = 0.23 deg^2

Lies, damn lies and statistics

Example 2

BBC NEWS

UK POLITICS

17 August 2011 Last updated at 11:31 GMT

'Worrying' jobless rise needs urgent action - Labour

Labour have said ministers need to take urgent action to reverse the "very worrying" rise in unemployment.

The number of people out of work rose by 38,000 to 2.49 million in the three months to June, official figures show.

Why was this poor statistics? The conclusion was not statistically significant because of the sampling errors. The "official figures" also showed that the 95% confidence interval was -49,000 to +125,000

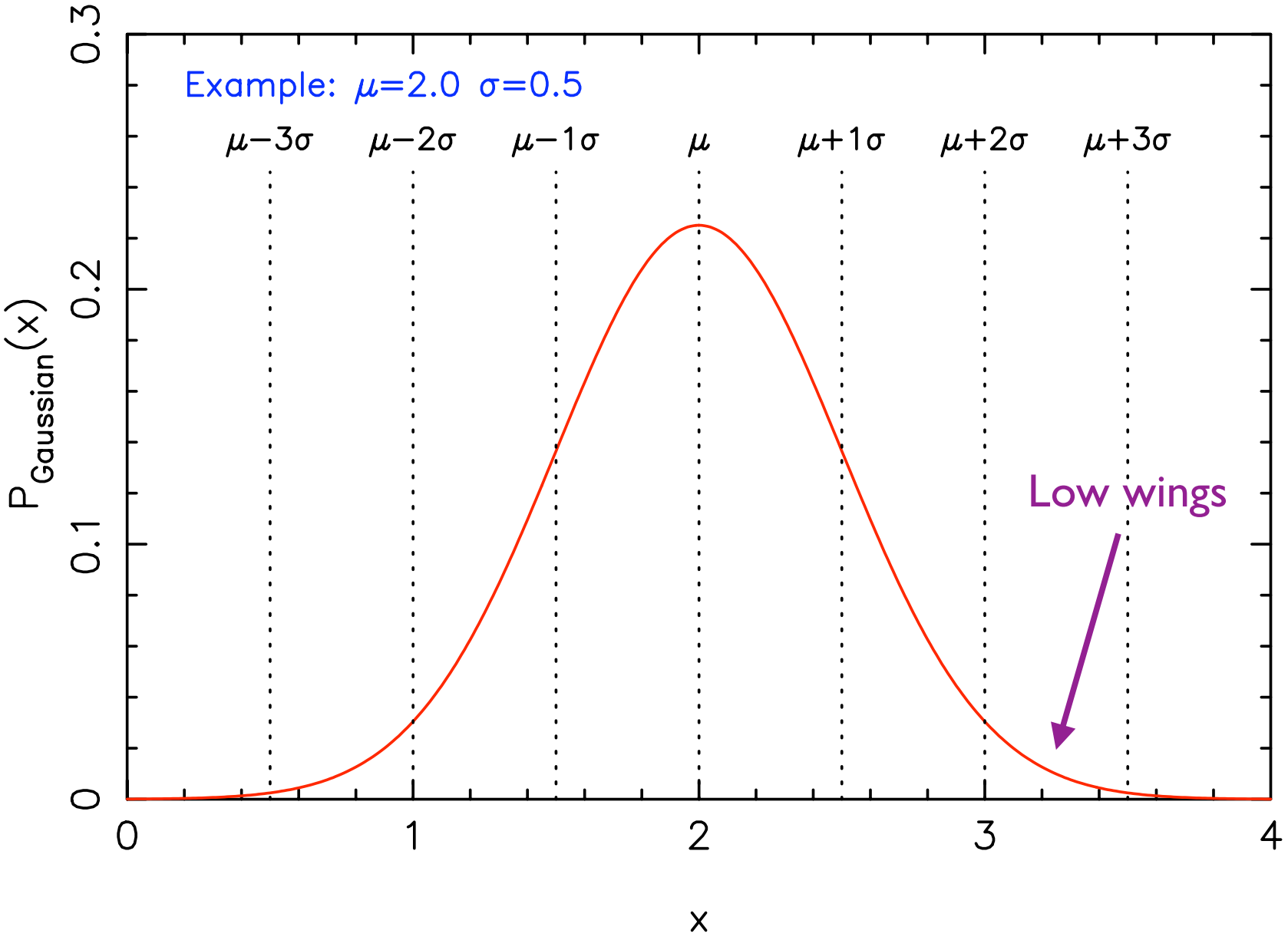
Gaussian distribution

$$P_{\text{Gaussian}}(x) = \frac{\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}{\sigma\sqrt{2\pi}}$$

x = continuous variable μ = mean σ = standard deviation

- Why is the Gaussian or Normal distribution such a ubiquitous and important probability distribution?
- High-N limit for binomial and Poisson distributions
- Central limit theorem : if we average together variables drawn many times from any probability distribution, the resulting averages will follow a Gaussian!

Gaussian distribution



"N-sigma confidence"

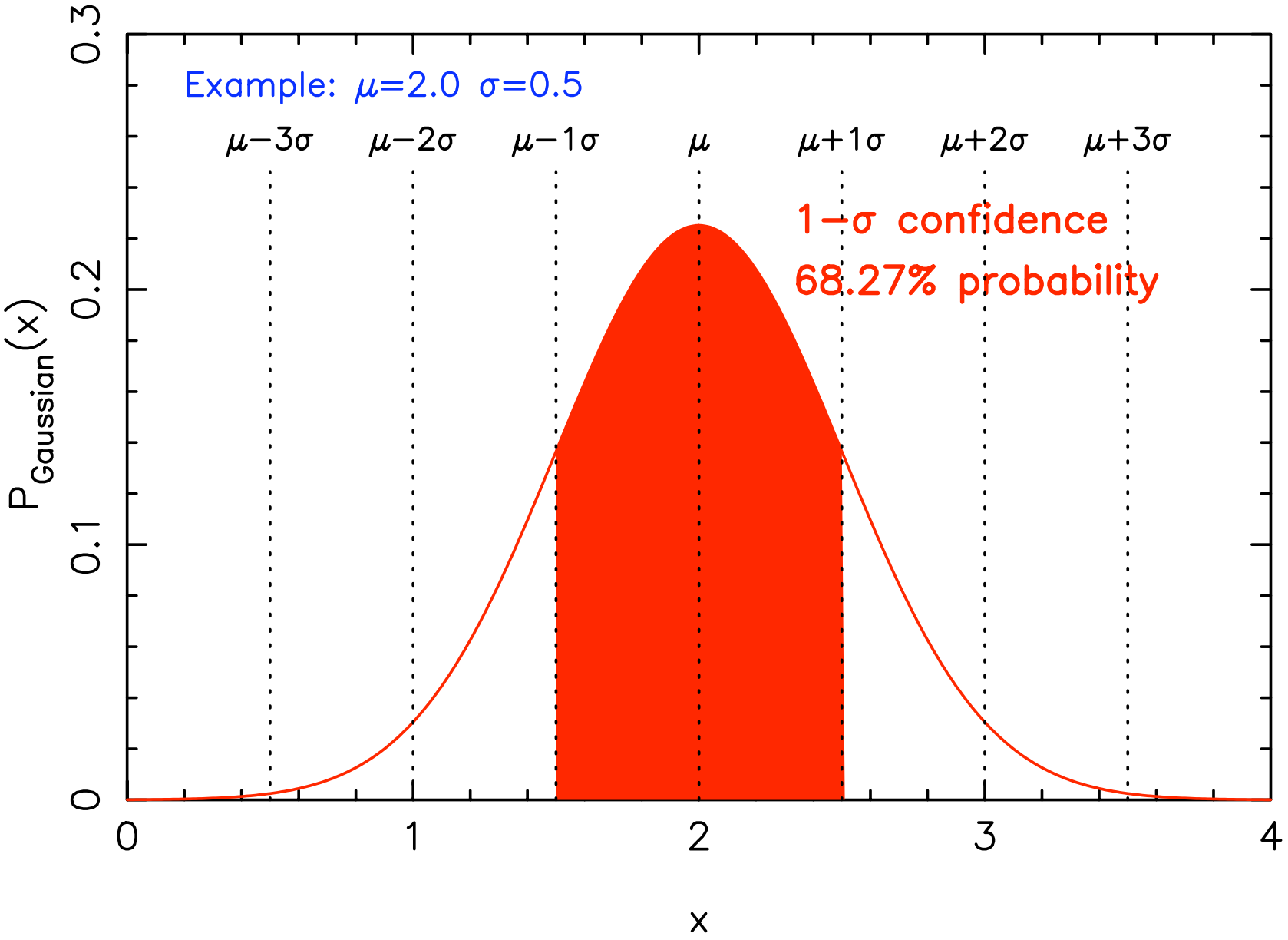
$$P_{\text{Gaussian}}(x) = \frac{\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}{\sigma\sqrt{2\pi}}$$

x = continuous variable μ = mean σ = standard deviation

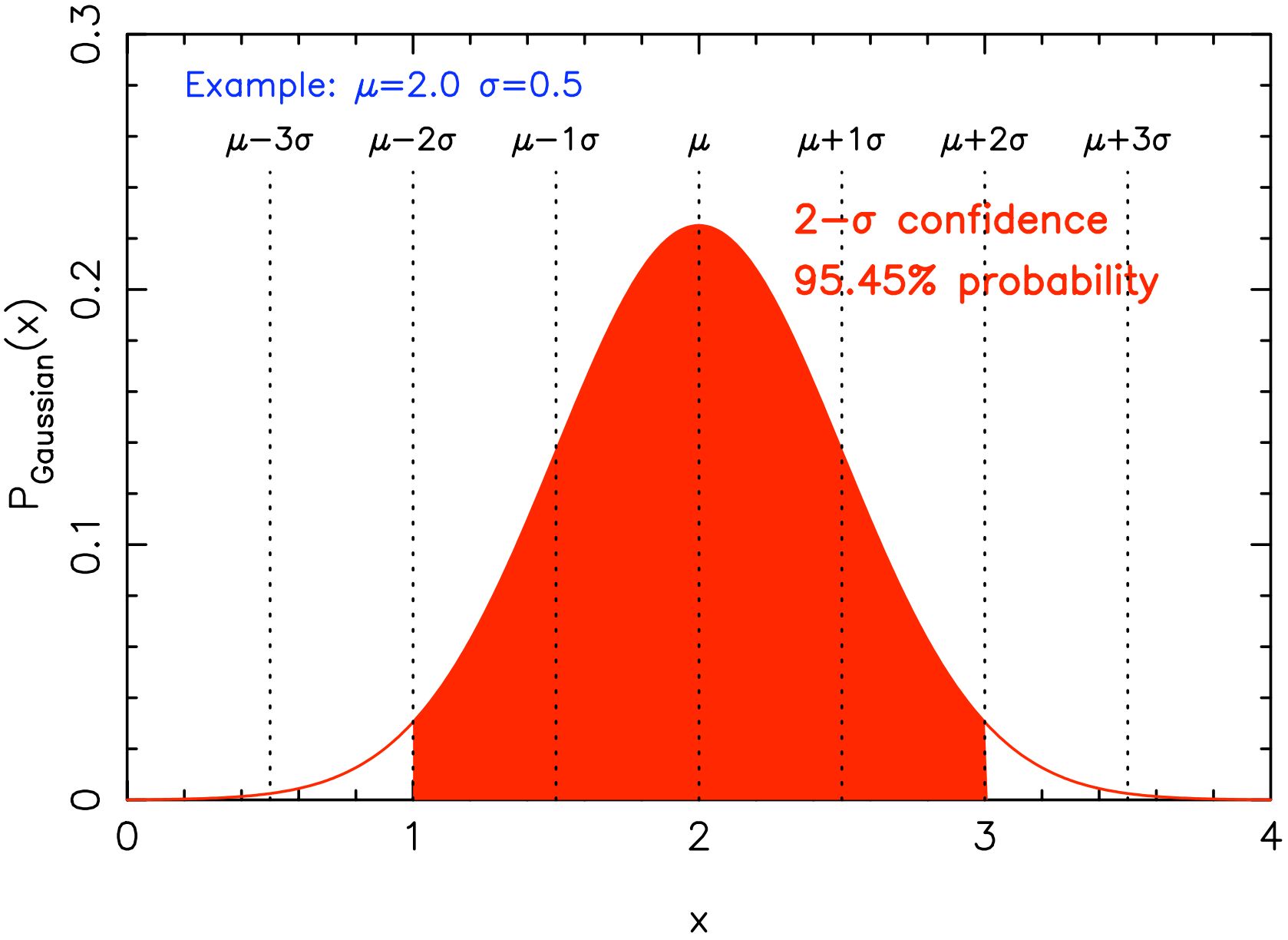
- Probability contained with +/- 1,2,3 standard deviations is (68.27, 95.45, 99.73)%

- For example, if a statement is said to have been verified with 3-sigma confidence, the implication is that it is expected to be true with a probability of 99.73%

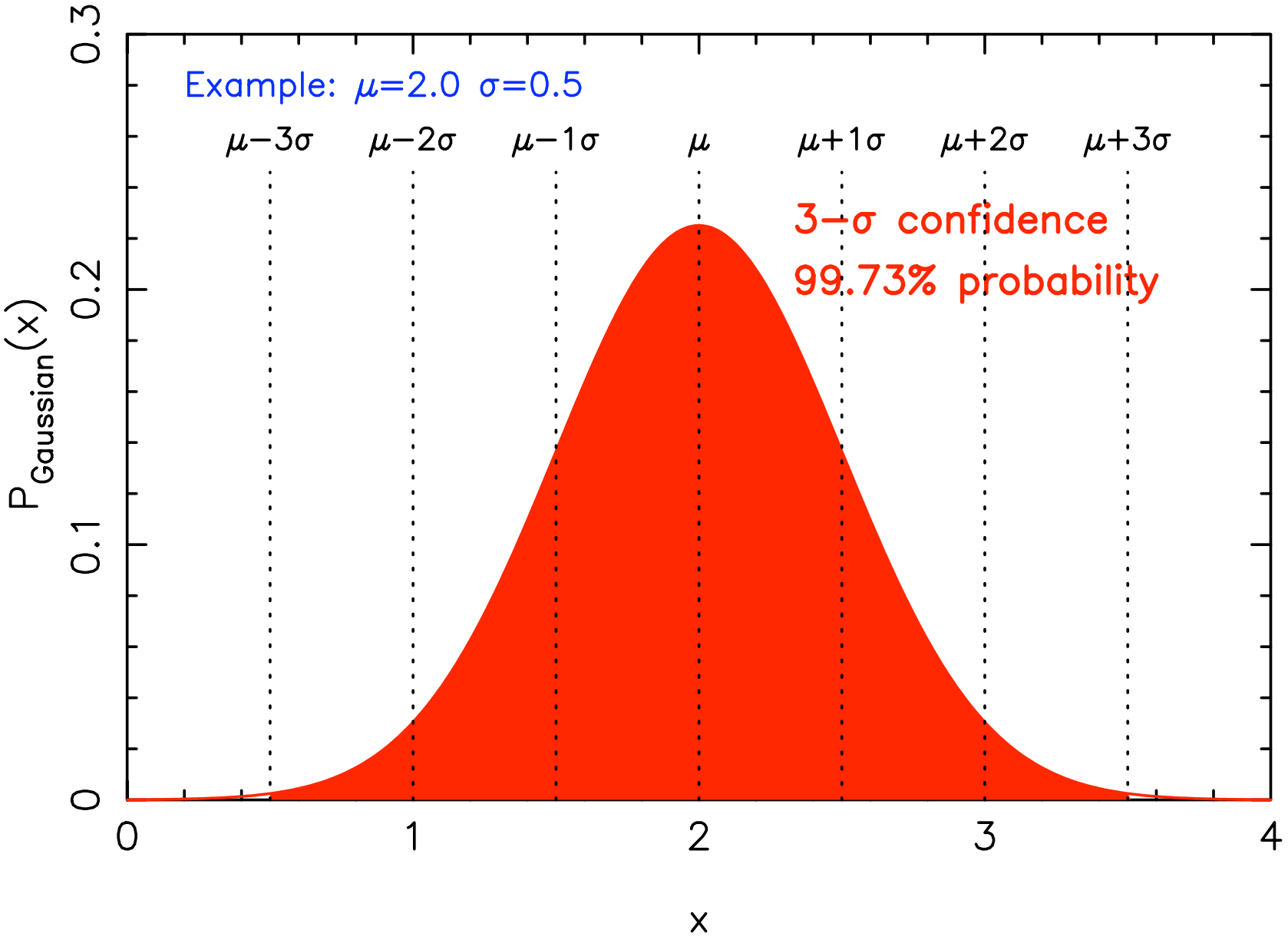
Gaussian distribution



Gaussian distribution



Gaussian distribution



Error propagation

- We have measurements and errors of some variables. What is the error in a function of those variables?

- Linear function of variables (x,y) with constants (a,b) :

$$z = a\,x + b\,y$$

$$\mathrm{Var}(z) = a^2\,\mathrm{Var}(x) + b^2\,\mathrm{Var}(y)$$

Error propagation

- Non-linear function of one variable x :

$$z = f(x)$$

$$\sigma_z = \left|\frac{\partial f}{\partial x}\right| \sigma_x$$

- Non-linear function of two variables x and y :

$$z = f(x, y)$$

$$\mathrm{Var}(z) = \left(\frac{\partial f}{\partial x}\right)^2 \mathrm{Var}(x) + \left(\frac{\partial f}{\partial y}\right)^2 \mathrm{Var}(y)$$

- Small print : These are approximations which assume the derivatives are constant. Will fail badly in some cases, e.g. z=x/y when y~0 !

Error propagation

- Example 6 : a galaxy of absolute magnitude M = -20 is observed to have apparent magnitude m = 20.0 +/- 0.2. What is the galaxy luminosity distance D_L and its error, assuming $m - M = 5 \log_{10}(D_L) + 25$?

- $D_L = 10^{0.2(m-M-25)} = 1000$ Mpc

- Error : if $y = 10^x$ then $dy/dx = y \log_e(10)$

- $sig(D_L) = D_L \log_e(10)\, sig(m) = 461$ Mpc [asymmetric error]

- Compare with exact range : (631-1585)/2 = 477 Mpc

Error propagation

- Example 7 : the total mass of a binary star system in solar masses is $M = a^3/P^2$ where a=mean separation in A.U. and P=period in years. For alpha Centauri, a = 23.7 +/- 1.0 A.U. and P = 79.9 +/- 1.0 years. What is the total mass of the system and its error?

- $sig_M^2 = (9a^4/P^4)\ sig_a^2 + (4a^6/P^6)\ sig_P^2$

- $(sig_M/M)^2 = 9\ (sig_a/a)^2 + 4\ (sig_P/P)^2$

- M = 2.08 +/- 0.27 solar masses

Optimal combination of data

- We have N independent estimates x_i of some quantity, with varying errors. What is our best combined estimate?

- A simple average? $$y = \frac{\sum_{i=1}^{N} x_i}{N}$$

- Not optimal because we want to give more weight to the more precise estimates. An unbiased estimator is :

$$y = \frac{\sum_{i=1}^{N} w_i \, x_i}{\sum_{i=1}^{N} w_i}$$

Optimal combination of data

- The combined error is minimized for inverse-variance weighting :

$$y = \frac{\sum_{i=1}^{N} x_i/\sigma_i^2}{\sum_{i=1}^{N} 1/\sigma_i^2}$$

- In this case :

$$\frac{1}{\mathrm{Var}(y)} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}$$

- Ensure that the data you are combining is self-consistent (i.e., systematic errors are not dominant)

Optimal combination of data

- Example 8 : we have N=5 measurements of a quantity : (7.4 +/- 2.0, 6.5 +/- 1.1, 4.3 +/- 1.7, 5.5 +/- 0.8, 6.0 +/- 2.5). What is the optimal estimate of this quantity and the error in that estimate? A further measurement 3.0 +/- 0.2 is added. How should our estimate change?

- Estimate = 5.81 +/- 0.56
  [Compare unweighted estimate = 5.94 +/- 0.77]

- The revised estimate would be 3.31 +/- 0.19, but the new measurement is an outlier and further investigation is needed