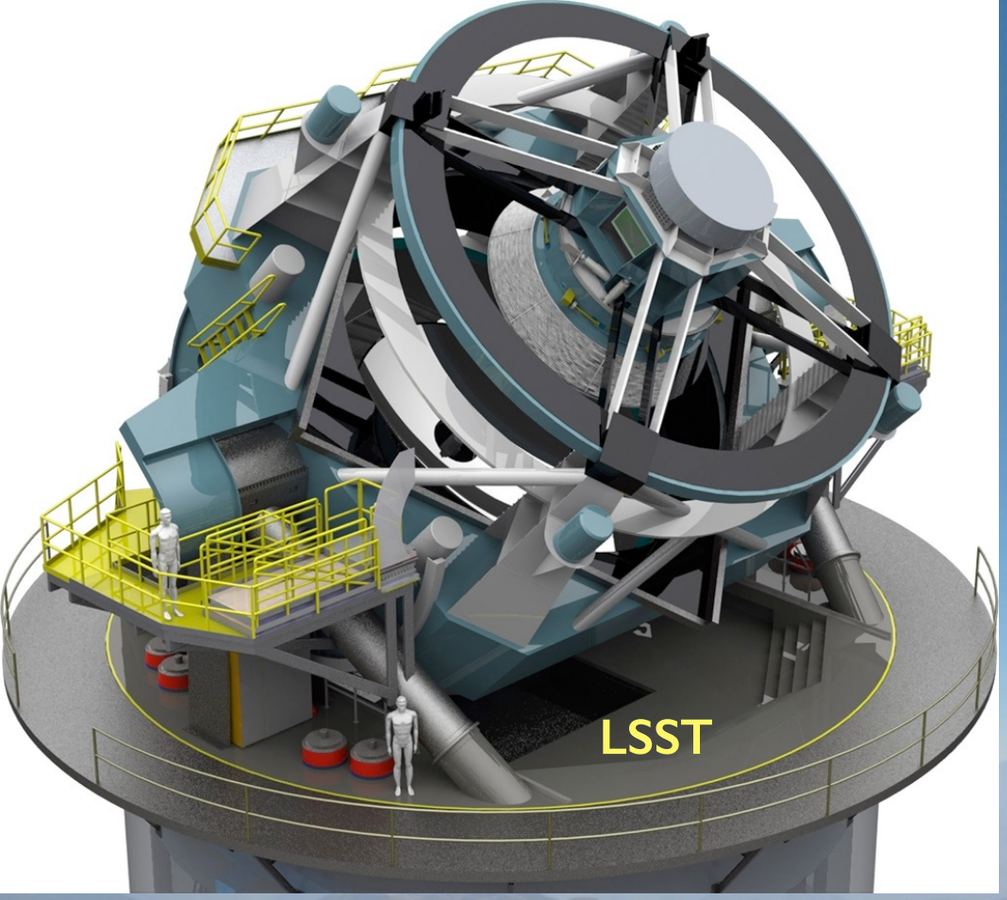
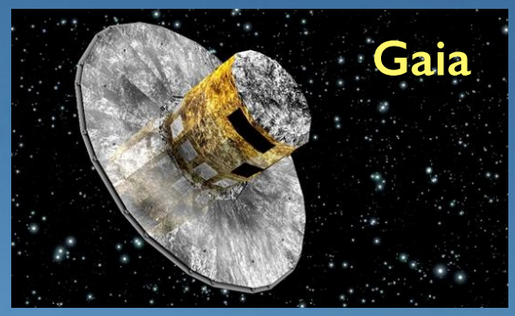
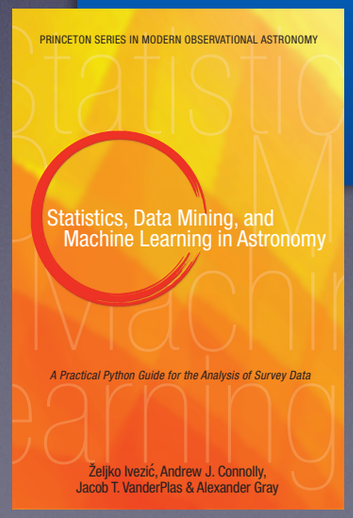




Brief Introduction to Statistics

Astr 323, Lecture 10

Spring 2014, University of Washington



Outline

- **How to estimate location and scale?**
  - Mean, median, std. dev. and their errors
  - Central Limit Theorem
  - Robust statistics
  - Gaussianity, Chi-squared, and non-gaussianity
- **How to make a histogram?**
- **If we (only) had more time…**

• How to compute an "average" value?

o First, "average" can mean different quantities, most often the mean and median (often, and erroneously, "average" and "mean" are considered synonymous)

o Given a list of numbers, x_i, i=1…N, their mean is

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

o "Everyone" knows that! We often take the mean of many measurements to improve the accuracy of the final result.

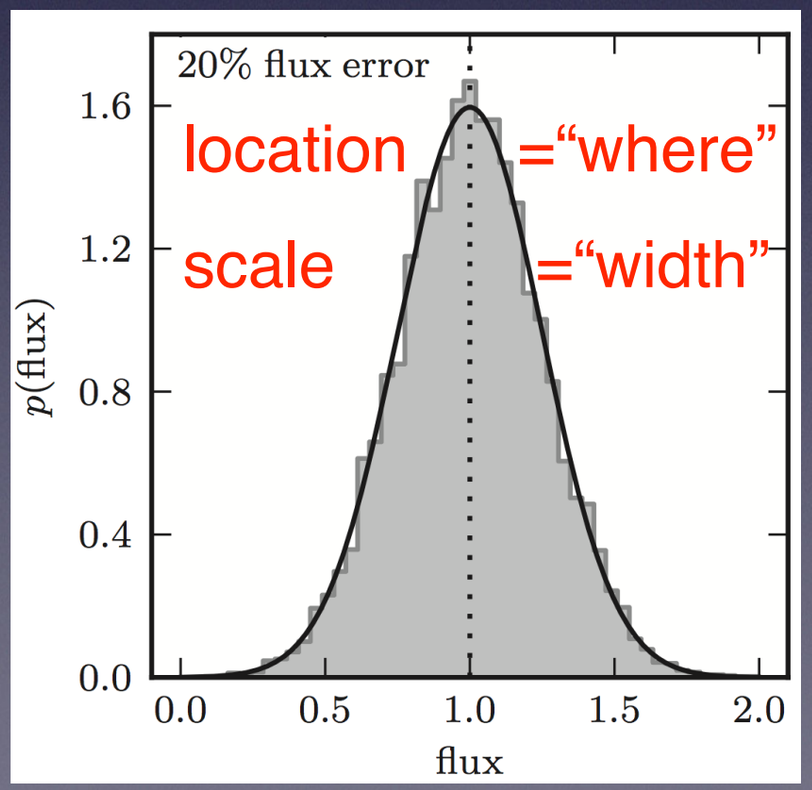
o What is not known by everyone is why and when this "averaging" works, and how to estimate how good it is. These answers are provided by statistics.

• How to estimate location and scale?

o A significant fraction of statistics is about using a list of numbers, x_i, i=1…N, drawn from some unknown distribution function, h(x), to estimate the properties of h(x). Here h(x) is a probability density function (pdf)

o In general, this one-dimensional case can be generalized to many dimensions, but here we'll keep it simple.

o First, let's see how we can quantify "the properties of h(x)"



This is an example of a Gaussian distribution: its location is 1.0 and it's scale is 0.2

Task: given a sample x_i, i=1…N, find the location and scale of the underlying (here Gaussian) distribution

• The most important h(x): Gaussian

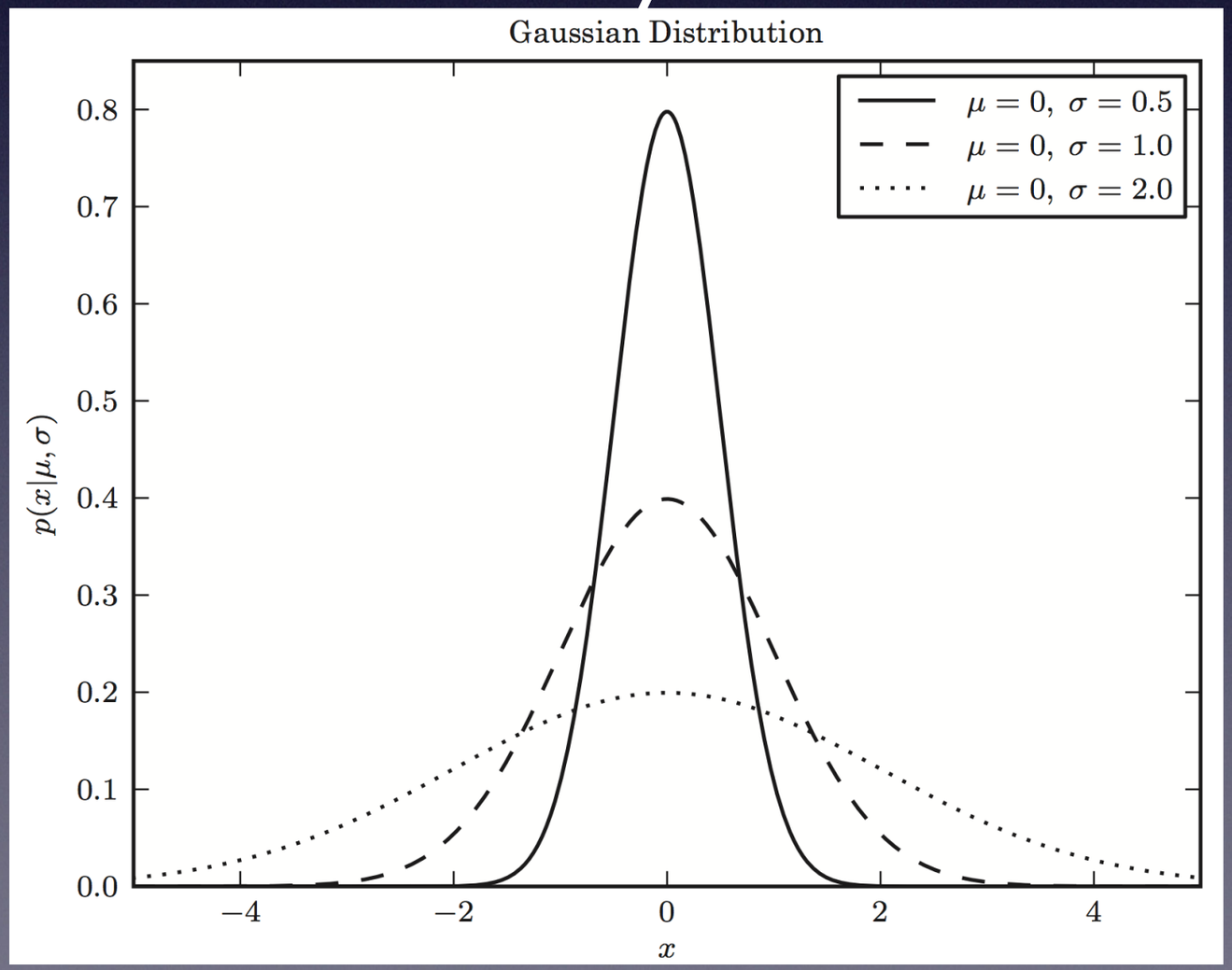
o Gaussian distribution is described by

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

a.k.a. Normal Distribution

$$\mathcal{N}(\mu,\sigma)$$

where | is pronounced "given". So, "given location parameter μ and scale parameter σ", p(x | μ, σ) gives the probability that a randomly drawn value will be between x and x+dx.



The integral of p over all possible values of x is unity and evaluated using the "Gauss error function" (which is not analytic):

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\left(-t^2\right) dt$$

• Why is Gaussian the most important h(x)?

o Because of the Central Limit Theorem:

Given an *arbitrary* distribution $h(x)$, characterized by its location μ and scale σ, the mean of N values x_i drawn from that distribution will approximately follow a Gaussian distribution with $N(\mu,\sigma/\sqrt{N})$, with the approximation accuracy improving with N.

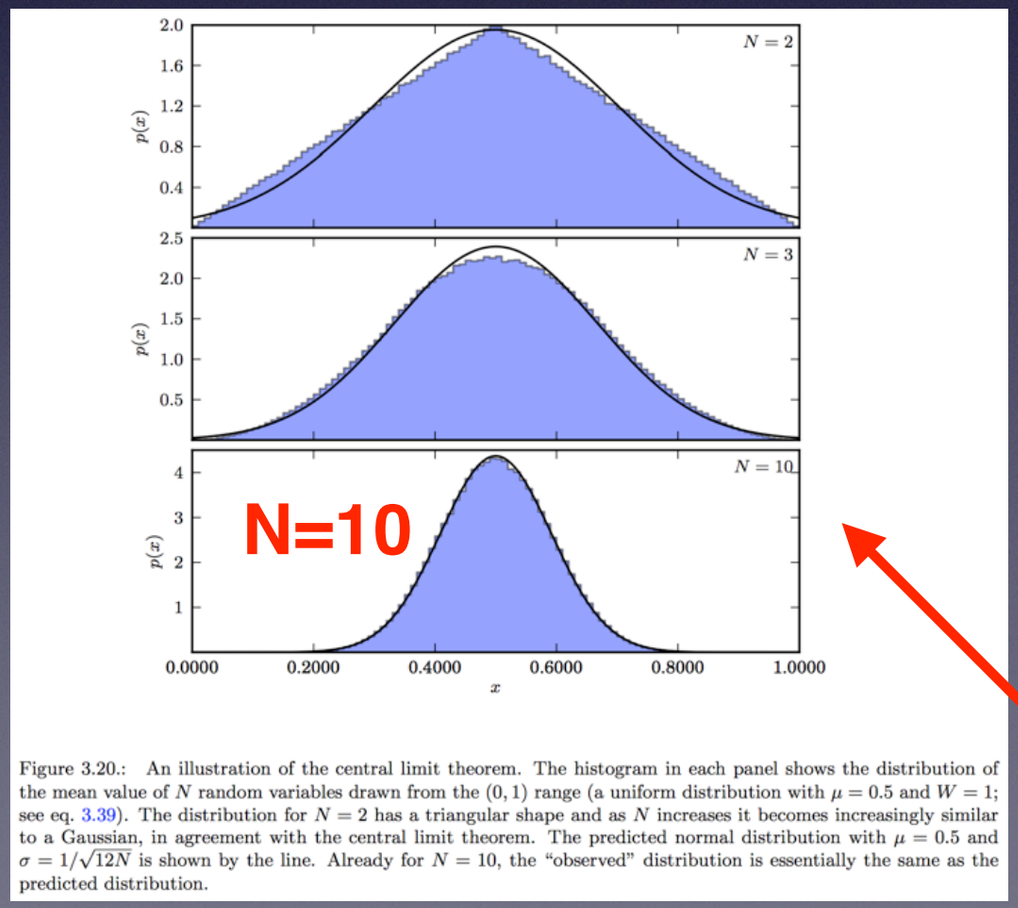
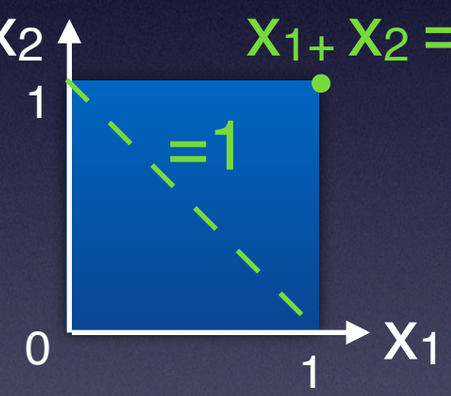


Figure 3.20.: An illustration of the central limit theorem. The histogram in each panel shows the distribution of the mean value of N random variables drawn from the $(0, 1)$ range (a uniform distribution with $\mu = 0.5$ and $W = 1$; see eq. 3.39). The distribution for $N = 2$ has a triangular shape and as N increases it becomes increasingly similar to a Gaussian, in agreement with the central limit theorem. The predicted normal distribution with $\mu = 0.5$ and $\sigma = 1/\sqrt{12N}$ is shown by the line. Already for $N = 10$, the "observed" distribution is essentially the same as the predicted distribution.



The CLT can be easily proven using standard tools from statistics, such as characteristic functions and convolutions. Here is an example of CLT in action based on a uniform distribution.

- **Why is Gaussian the most important h(x)?**
  - Because of the Central Limit Theorem:

Given an *arbitrary* distribution $h(x)$, characterized by its location μ and scale σ, the mean of N values x_i drawn from that distribution will approximately follow a Gaussian distribution with $N(\mu,\sigma/\sqrt{N})$, with the approximation accuracy improving with N.

- This is a remarkable result since the details of the distribution $h(x)$ are not specified - we can "average" our measurements (i.e., compute their mean value) and expect the $1/\sqrt{N}$ improvement in accuracy *regardless of details in our measuring apparatus!*

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

But note that it was implicitly assumed that h(x) has finite σ - not always true!
The Cauchy distribution: σ is undefined

• How we can quantify h(x)?

- Arithmetic mean (also known as the expectation value),

$$\mu = E(x) = \int_{-\infty}^{\infty} x h(x)\, dx$$

- Variance,

$$V = \int_{-\infty}^{\infty} (x-\mu)^2 h(x)\, dx$$

- Standard deviation,

$$\sigma = \sqrt{V}$$

- Skewness,

$$\Sigma = \int_{-\infty}^{\infty} \left(\frac{x-\mu}{\sigma}\right)^3 h(x)\, dx$$

- Kurtosis,

$$K = \int_{-\infty}^{\infty} \left(\frac{x-\mu}{\sigma}\right)^4 h(x)\, dx - 3$$

- p% quantiles (p is called a percentile), q_p,

$$\frac{p}{100} = \int_{-\infty}^{q_p} h(x)\, dx$$

Location parameter

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

What is σ for the Cauchy distr. above?

Scale parameter

Shape parameters

Parameters describing cumulative distribution

Location, scale, shape, and other parameters defined for h(x) can also be computed for a sample drawn from h(x).

In general, when estimating the above quantities for a sample of N measurements, the integral $\int_{-\infty}^{\infty} g(x)h(x)\,dx$ becomes proportional to the sum $\sum_i^N g(x_i)$, with the constant of proportionality $\sim (1/N)$. For example, the *sample arithmetic mean*, $\overline{x}$, and the *sample standard deviation*, s, can be computed via standard formulas,

When using h(x), these parameters are called population statistics, and when determined from data x_i, i=1…N, they are called sample statistics.

For example, location parameter μ for h(x) is estimated using the mean:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

and scale parameter σ for h(x) is estimated using the standard deviation:

(I know, blame statisticians, not me!)

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

Note: when h(x) is estimated from data, a different symbol should be used, e.g. f(x)

Location parameter μ for h(x) is estimated using the mean:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

and scale parameter σ is estimated using the standard deviation:

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

These estimators will NOT be exactly equal to μ and σ!
Each will be scattered around the true values (μ and σ) approximately following Gaussian distributions with the widths (scale parameters) given by:

the standard error of the mean

$$\sigma_{\overline{x}} = \frac{s}{\sqrt{N}}$$

often called "error bar"!

error of the standard deviation estimate s:

$$\sigma_s = \frac{s}{\sqrt{2(N-1)}}$$

So, given x_i, i=1…N, we can compute

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \pm \quad \sigma_{\overline{x}} = \frac{s}{\sqrt{N}}$$

the sample mean

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2} \quad \pm \quad \sigma_s = \frac{s}{\sqrt{2(N-1)}}$$

the sample standard deviation

What exactly did we compute?

In the majority of practical cases, x_i represent our N measurements of some fixed well-defined quantity x, and our inference about its value is summarized by the sample mean and the standard error of the mean, and assumed Gaussian distribution.

What about standard deviation?

So, given x_i, i=1…N, we can compute

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2} \pm \sigma_s = \frac{s}{\sqrt{2(N-1)}}$$

the sample standard deviation

What about standard deviation?

There are two limiting, but often realistic, cases in practice:

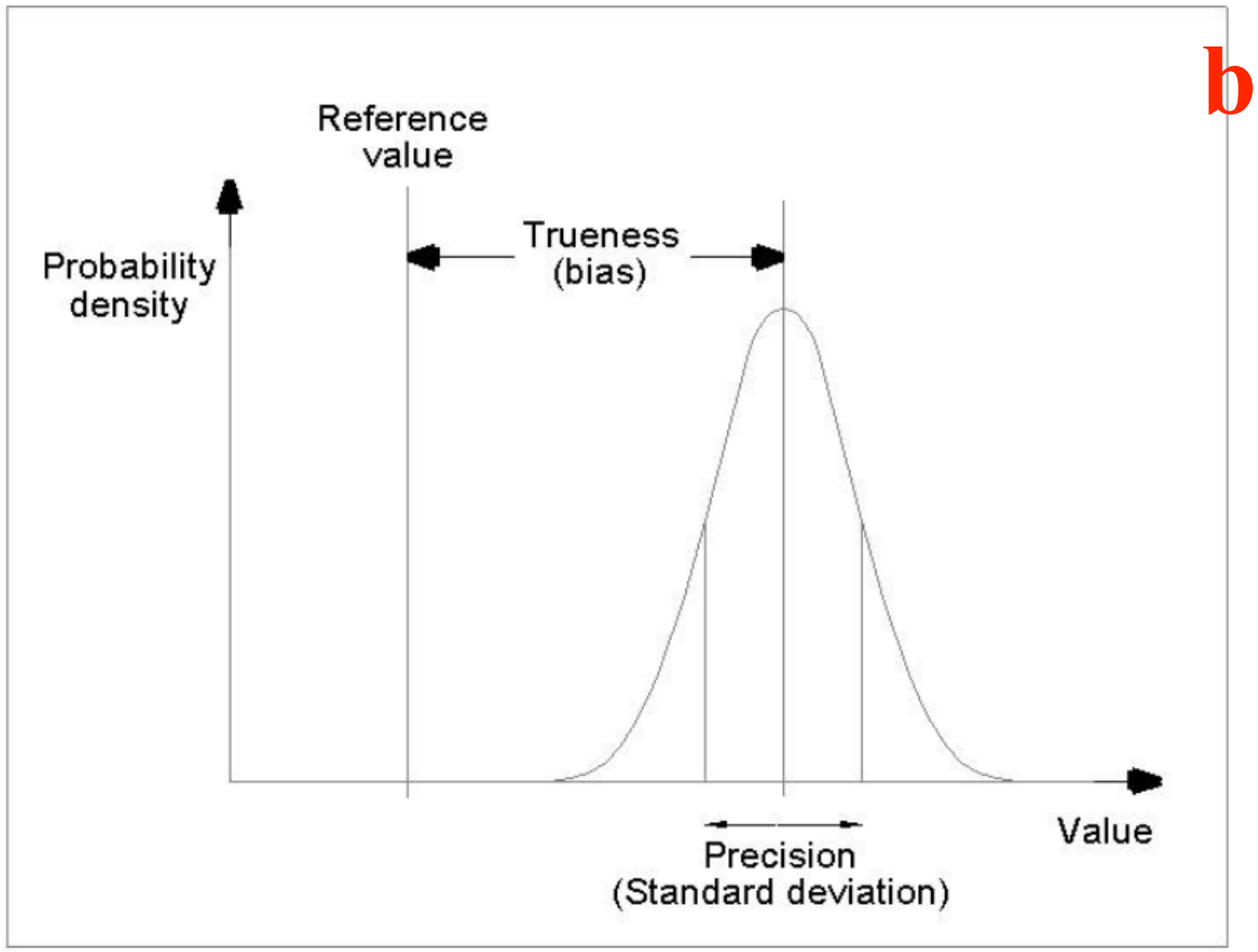
1) Our measurements have some fixed (a priori unknown) measurement error σ, and (s, σ_s) summarizes our knowledge of it

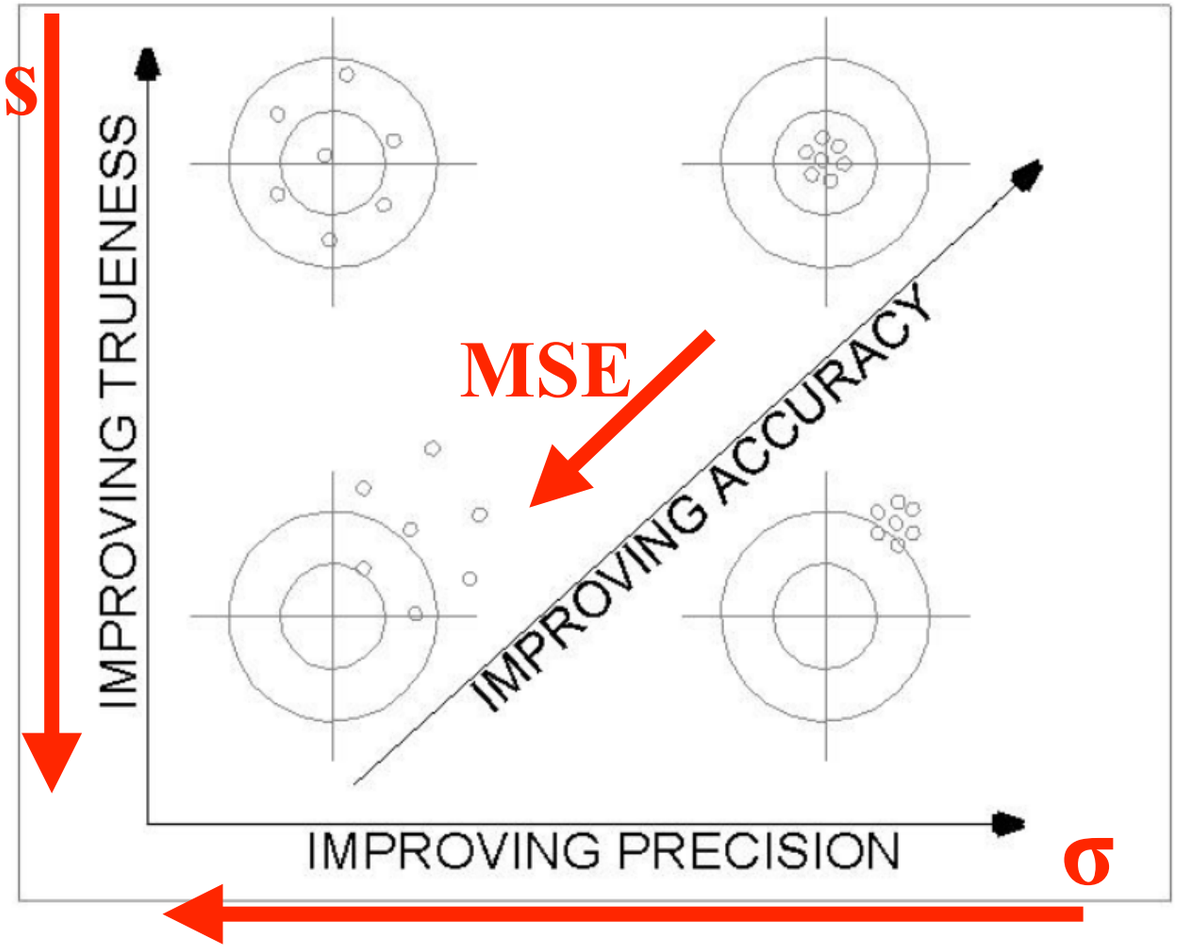
2) We measured quantity x which has its own intrinsic scatter σ (e.g. weight of a loaf of bread), and (s, σ_s) summarizes our knowledge of it (assuming here that the measurement error is negligible compared to σ).

Let's discuss case 1) in more detail…

• Measurements with known errors

Let's assume that we have N measurements x_i, and that for each measurement with know the corresponding error distribution, that is, the expected distribution of x_i around the true value μ (which we want to estimate)





Mean Squared Error: $\mathrm{MSE} = V + \mathrm{bias}^2$ (V=variance=σ^2)

• Measurements with known errors

Let's assume that we have N measurements x_i, and that for each measurement with know the corresponding error distribution, that is, the expected distribution of x_i around the true value μ (which we want to estimate)

Let's also assume that this known error distribution is given by Gaussian distribution, N(bias, σ_i). We will also assume here that bias = 0. (later we will relax these assumptions)

When all σ_i are same, they are called homoscedastic errors.
When σ_i vary, they are called heteroscedastic errors.

Task: given measurements (x_i, σ_i), i=1…N, find the best estimate of μ and its uncertainty

• Measurements with known errors

Task: given measurements (x_i, σ_i), i=1…N, find the best estimate of μ, let's call it μ^0,and its uncertainty, σ_μ

This problem problem can be solved using the Maximum Likelihood method, giving ("weighted" mean)

$$\mu^0 = \frac{\sum_i^N w_i x_i}{\sum_i^N w_i}$$

with weights $w_i = \sigma_i^{-2}$

the uncertainty of μ^0 is

$$\sigma_\mu = \left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \right)^{-1/2} = \left(\sum_{i=1}^{N} w_i \right)^{-1/2}$$

Note that in case of homoscedastic errors (all $\sigma_i = \sigma$), μ^0 simply becomes arithmetic mean that we introduced earlier.

• Measurements with known errors

Since the known error distribution is supposed to follow Gaussian distribution, that is, each x_i is drawn from $N(\mu,\sigma_i)$, the distribution of the quantity

$$z_i = (x_i - \mu) / \sigma_i$$

must be given by the Gaussian N(0,1).

If z_i is NOT distributed as N(0,1), then we have a problem: either error distribution is not Gaussian, or the values of σ_i are unreliable, or the underlying quantity does not have a fixed value μ.

To test whether z_i is distributed as N(0,1), we essentially test whether its mean is consistent with 0, and whether its standard deviation is consistent with 1. In practice, we use…

• Chi-squared distribution

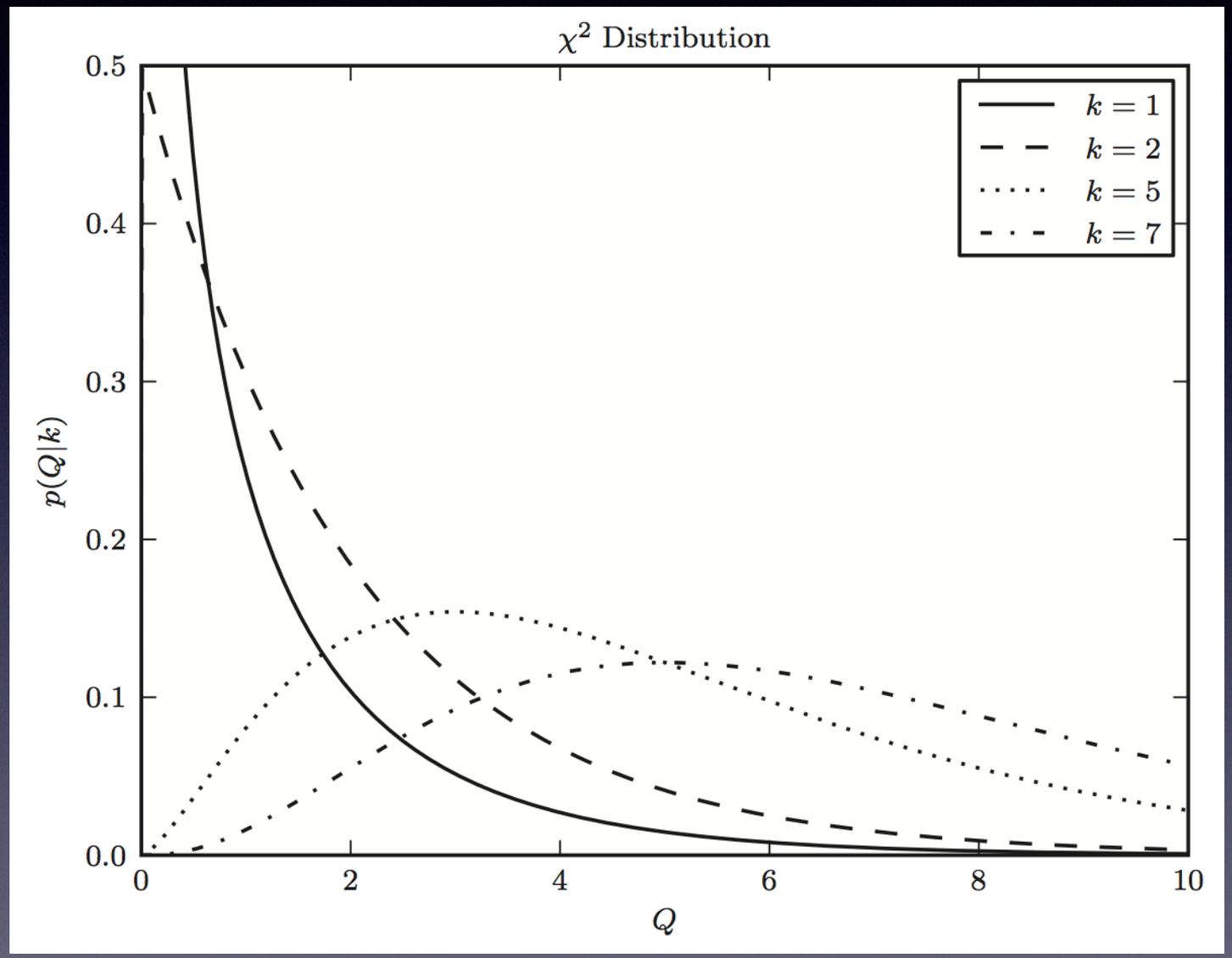
We compute quantity where $z_i = (x_i - \mu) / \sigma_i$

$$Q = \sum_{i=1}^{N} z_i^2$$

The important result here is that in case of Gaussian x_i, Q is distributed as the chi-squared distribution:

with k=N degrees of freedom

Expectation value: k
Std. deviation: $\sqrt{(2k)}$



$$p(Q|k) \equiv \chi^2(Q|k) = \frac{1}{2^{k/2}\Gamma(k/2)} Q^{k/2-1} \exp(-Q/2) \quad \text{for } Q > 0$$

$\Gamma(k) = (k-1)!$

• Chi-squared distribution

We compute quantity where $z_i = (x_i - \mu) / \sigma_i$

$$Q = \sum_{i=1}^{N} z_i^2$$

For large N (say, >10 or so), the chi-squared distribution approximately morphs into good old Gaussian distribution $N(k, \sqrt{2k})$.

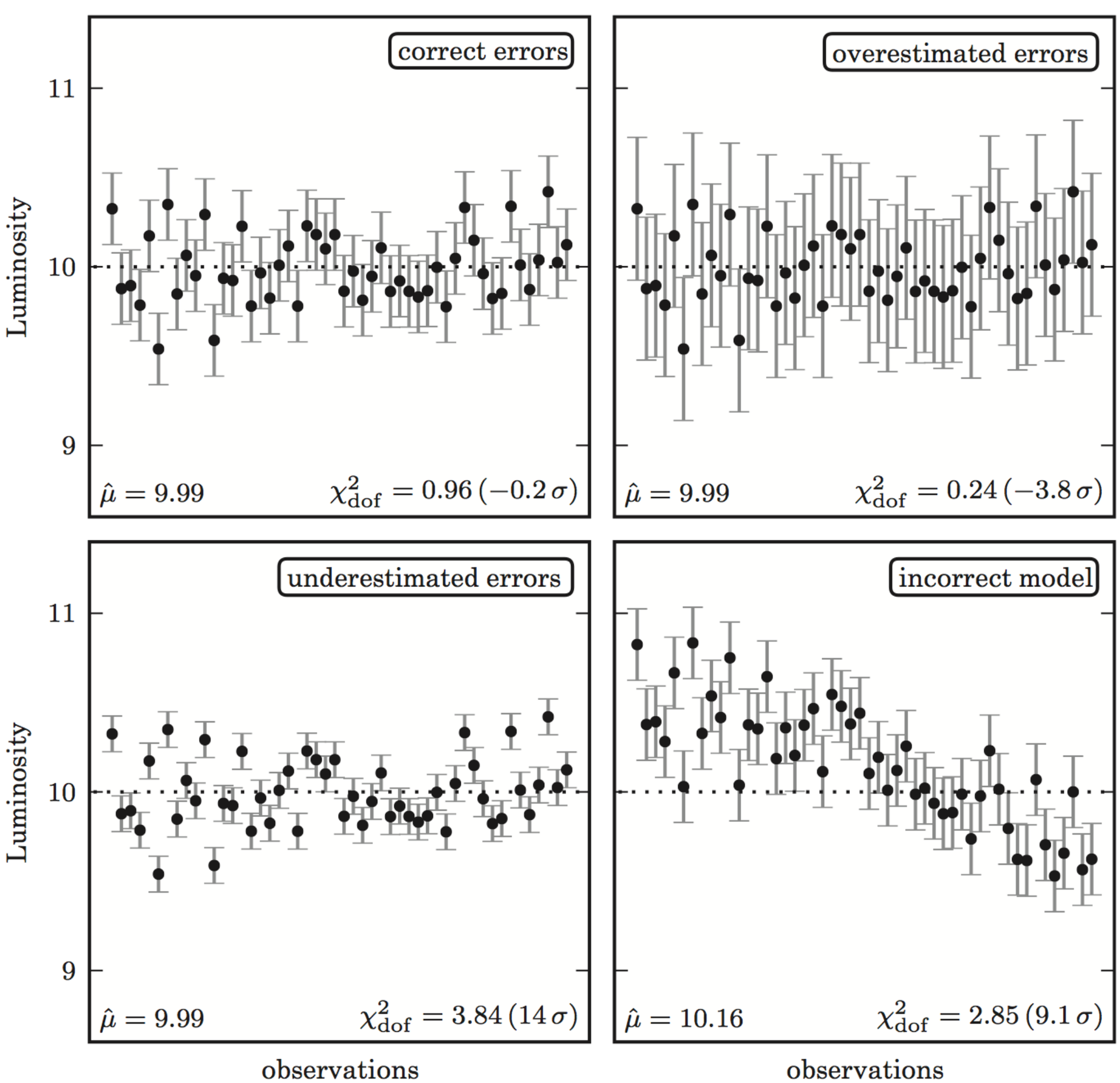
In practice, the "chi-squared per degree of freedom" is often used:

$$\chi^2_{dof} = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{x_i - \overline{x}}{\sigma_i} \right)^2$$

Each point deviates "on average" by one "error bar"

We expect χ^2_{dof} to be 1 to within a few $\sqrt{2/(N-1)}$

Note that χ^2_{dof} is essentially an estimate of the standard deviation squared for the quantity z_i


correct errors
overestimated errors
underestimated errors
incorrect model
Luminosity
11
10
9
$\hat{\mu} = 9.99$
$\chi^2_{\text{dof}} = 0.96\,(-0.2\,\sigma)$
$\hat{\mu} = 9.99$
$\chi^2_{\text{dof}} = 0.24\,(-3.8\,\sigma)$
$\hat{\mu} = 9.99$
$\chi^2_{\text{dof}} = 3.84\,(14\,\sigma)$
$\hat{\mu} = 10.16$
$\chi^2_{\text{dof}} = 2.85\,(9.1\,\sigma)$
observations
observations

• Chi-squared distribution

$$\chi^2_{dof} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{x_i - \overline{x}}{\sigma_i}\right)^2$$

We expect χ^2_{dof} to be 1 to within a few $\sqrt{2/(N-1)}$

If not, then we have a problem: either error distribution is not Gaussian, or the values of σ_i are unreliable, or the underlying quantity x does not have a fixed value μ.

If $(\chi^2_{dof}-1)$ is M times larger than $\sqrt{2/(N-1)}$, then we have M-sigma significant detection of the variability of x.

- **Robust statistics**

$$\chi^2_{dof} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{x_i - \overline{x}}{\sigma_i}\right)^2$$

χ^2_{dof} is very sensitive to "outliers", as is standard deviation.

If the error distribution is **not** Gaussian, the Central Limit Theorem tells us that the mean value will remain a good estimator of the location as long as the error distribution has a finite standard deviation.

Sometime distributions have infinite standard deviation, e.g. the Cauchy (Lorentzian) distribution:

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

- **Robust statistics**

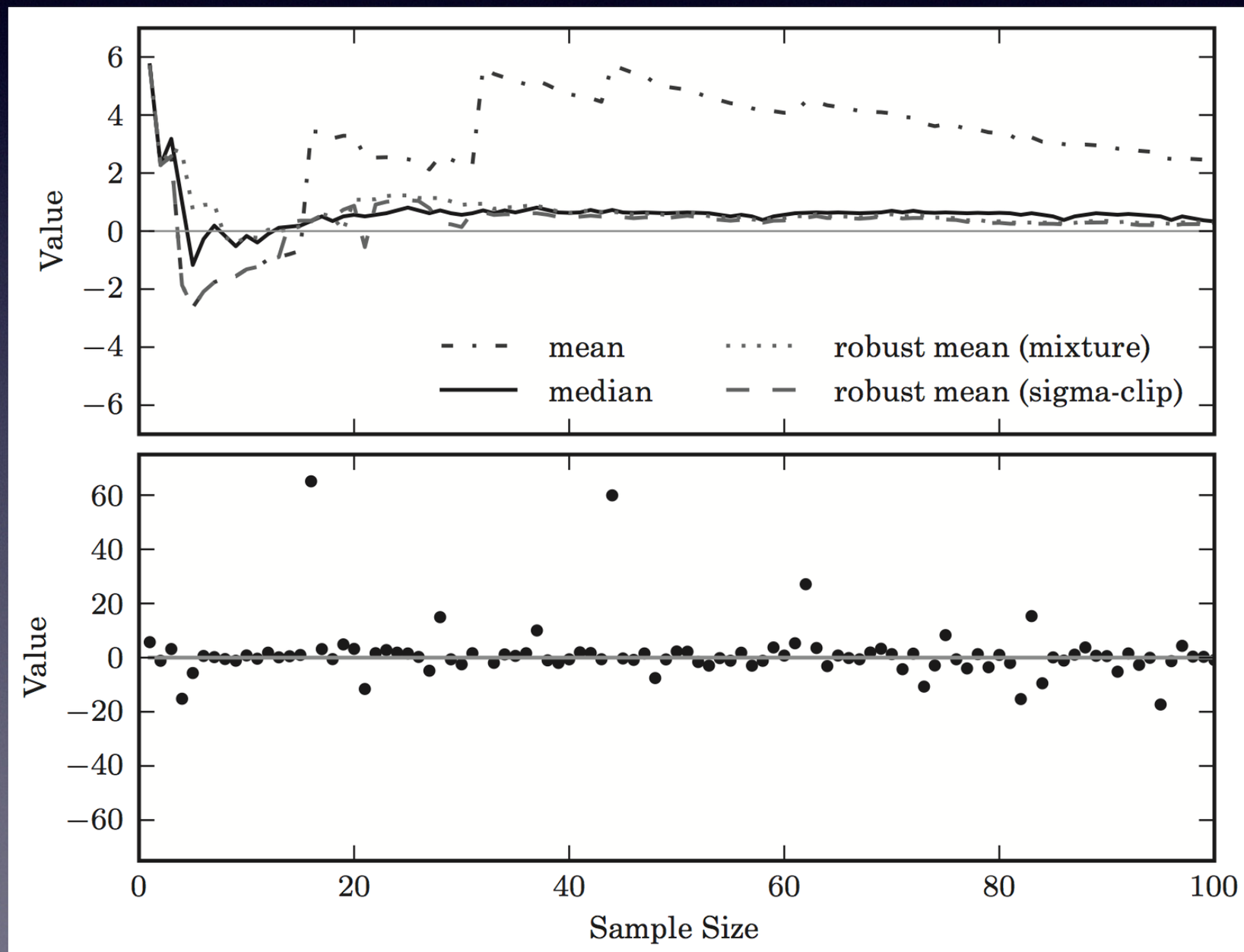
The Cauchy (Lorentzian) distribution:

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

Task: given measurements x_i, i=1…N, drawn from the Cauchy distribution, find the best estimate of μ, let's call it μ^0,and its uncertainty, σ_μ

In this case, using the mean value is a very bad idea!

Use the median instead!



- **Robust statistics**

Task: given measurements x_i, i=1…N, drawn from the Cauchy distribution, find the best estimate of μ, let's call it μ^0,and its uncertainty, σ_μ

Use the median value of x_i as an estimate of location, μ^0
The scale parameter ("width") can be estimated from the interquartile range (note: q_{50} is the median):

$$\sigma_G = 0.7413\,(q_{75} - q_{25})$$

In the case of Gaussian, σ_G is equal to standard deviation (σ)

The uncertainty of μ^0 (i.e. of the median) can be estimated from

$$\sigma_\mu = \sqrt{\frac{\pi}{2N}}\,\sigma_G$$

- **Robust statistics**

Median and σ_G are good estimators of location and scale parameters also in cases when outliers are present (e.g. “real data”)

The price we pay for using the median instead of the mean is 25% larger uncertainty for the former than for the latter (assuming nearly Gaussian distributions). This is often good price to pay to avoid catastrophic failures!

• Chi-squared – robust version

$$\chi^2_{dof} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{x_i - \overline{x}}{\sigma_i}\right)^2$$

We expect χ^2_{dof} to be 1 to within a few $\sqrt{2/(N-1)}$

If not, then we have a problem: either error distribution is not Gaussian, or the values of σ_i are unreliable, or the underlying quantity x does not have a fixed value μ.

Therefore, if the error distribution is suspected to have slow-falling tails, such as the Cauchy distribution, or outliers are present, use σ_G to estimate the width of z_i and thus robust χ^2_{dof}, instead of the formula above.

What is a histogram?

- **---> Data modeled by a step function**

Assuming that we have selected a bin size, Δ_b, the N values of x_i are sorted into M bins, with the count in each bin n_k, $k = 1, \ldots, M$. If we want to express the results as a properly normalized $f(x)$, with the values f_k in each bin, then it is customary to adopt

$$f_k = \frac{n_k}{\Delta_b N}. \tag{4.80}$$

The unit for f_k is the inverse of the unit for x_i.

Each estimate of f_k comes with some uncertainty. It is customary to assign "error bars" for each n_k equal to $\sqrt{n_k}$ and thus the uncertainty of f_k is

$$\sigma_k = \frac{\sqrt{n_k}}{\Delta_b N}. \tag{4.81}$$

This practice assumes that n_k are scattered around the true values in each bin (μ) according to a Gaussian distribution, and that error bars enclose the 68% confidence range for the true value. However, when counts are low this assumption of Gaussianity breaks down and the Poisson distribution should be used instead. For example, according to the Gaussian distribution, negative values of μ have nonvanishing probability for small n_k (if $n_k = 1$, this probability is 16%). This is clearly wrong since in counting experiments, $\mu \geq 0$. Indeed, if $n_k \geq 1$, then even $\mu = 0$ is clearly ruled out. Note also that $n_k = 0$ does not necessarily imply that $\mu = 0$: even if $\mu = 1$, counts will be zero in $1/e \approx 37\%$ of cases. Another problem is that the range $n_k \pm \sigma_k$ does not correspond to the 68% confidence interval for true μ when n_k is small. These issues are important when fitting

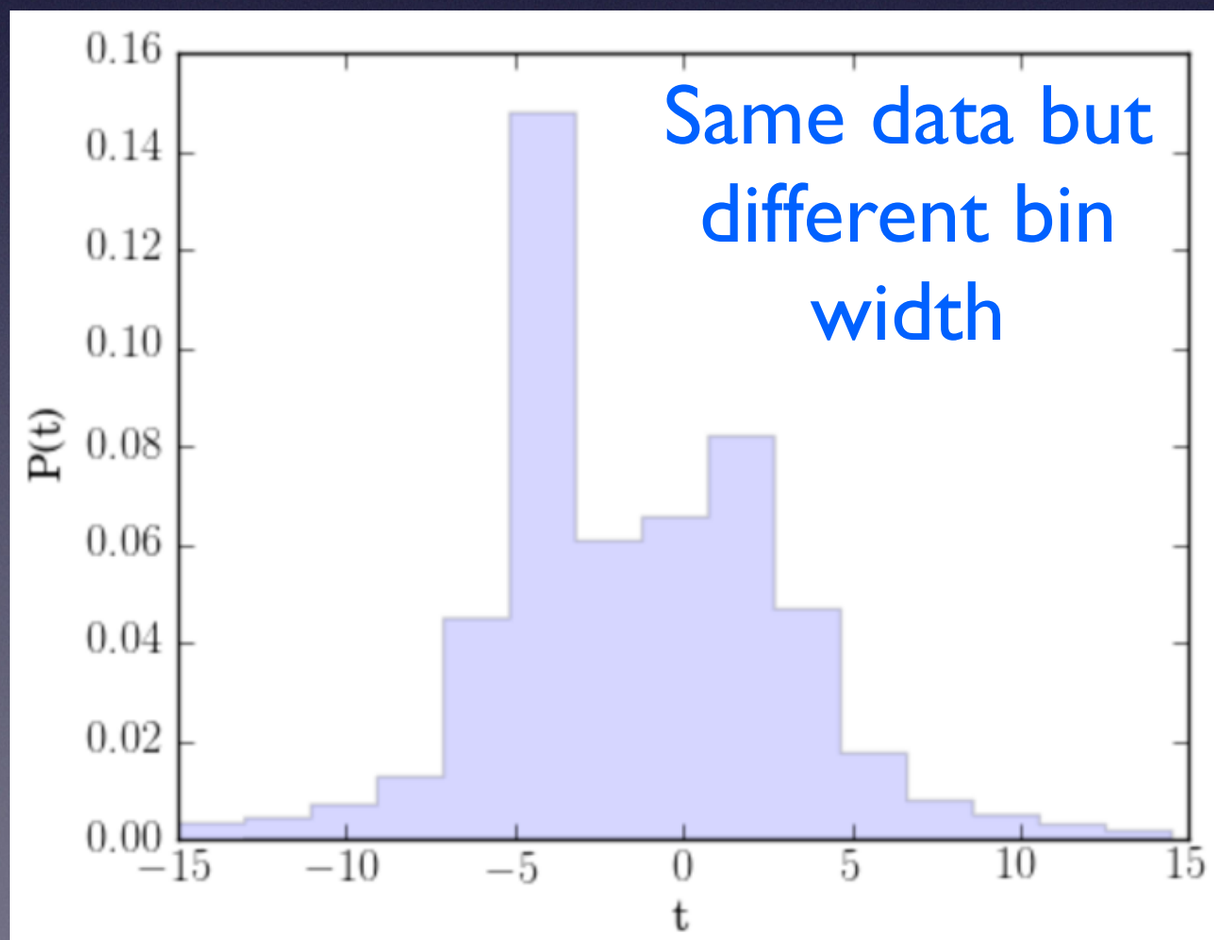
What is a histogram?

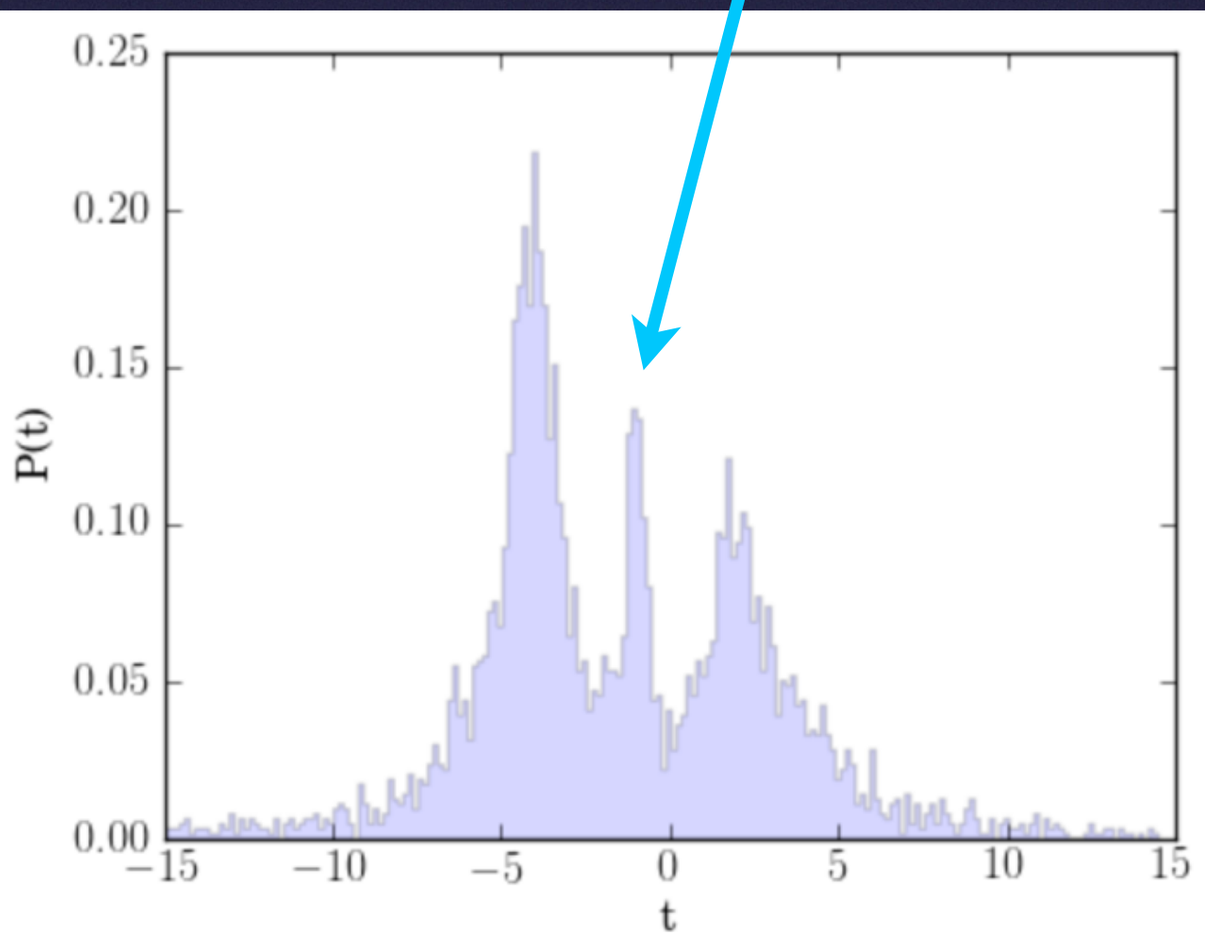
- ---> Data modeled by a step function
- How do we determine/estimate/guess the bin width?
- Do all the bins have to have the same width?
- Do we really have to bin data to estimate model parameters?

What is a histogram?

- ---> Data modeled by a step function
- How do we determine/estimate/guess the bin width?
- Do all the bins have to have the same width?
- Do we really have to bin data to estimate model parameters?

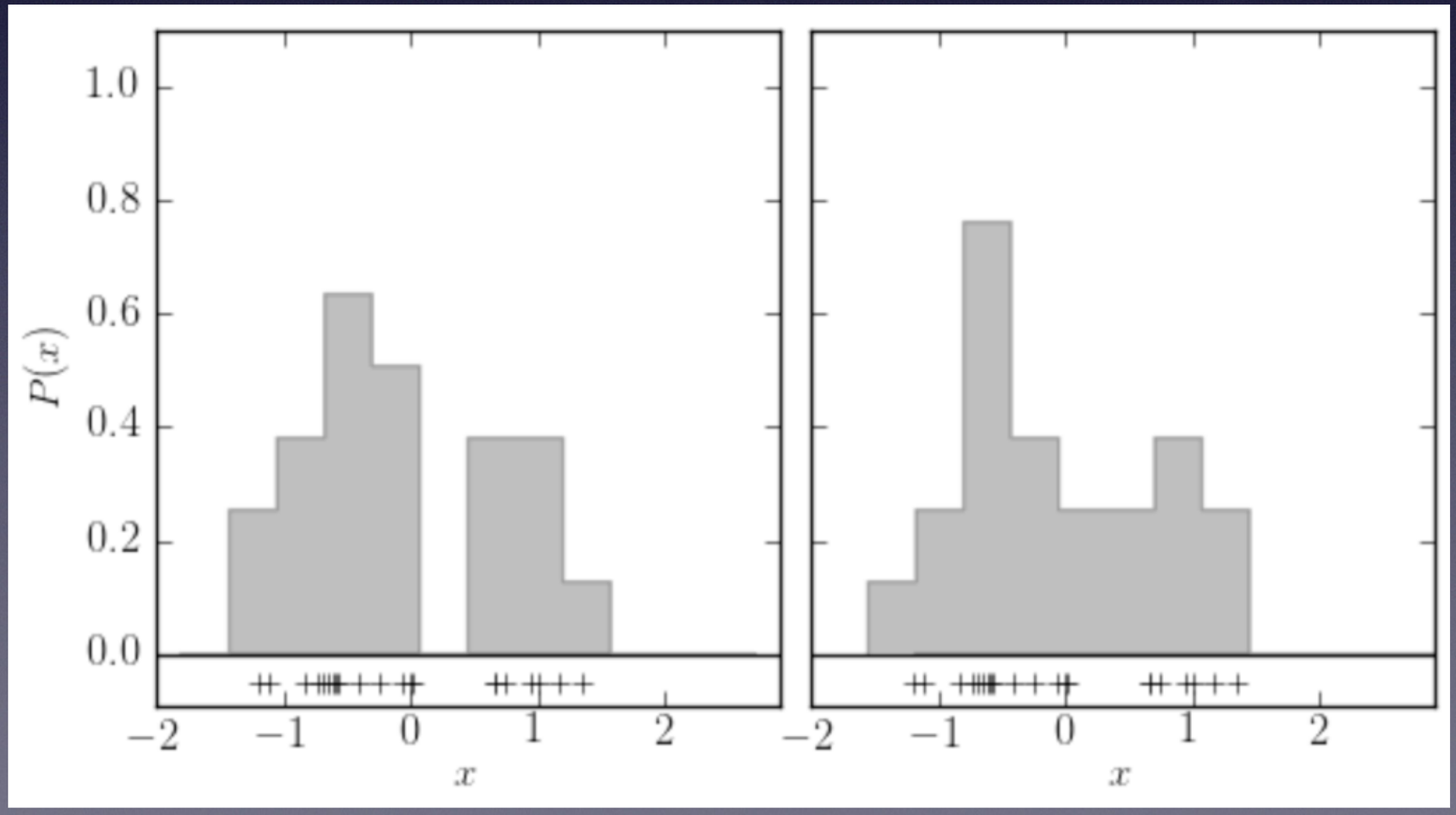
Should we believe the middle peak?





What is a histogram?

- ---> Data modeled by a step function
- How do we determine/estimate/guess the bin width?
- Do all the bins have to have the same width?
- Do we really have to bin data to estimate model parameters?



Despite the same bin width, a small offset in bin placement can give very different impressions about data behavior

Simple rules for estimating bin width

Various proposed methods for choosing optimal bin width typically suggest a value proportional to some estimate of the distribution's scale, and decreasing with the sample size. The most popular choice is "Scott's rule" which prescribes a bin width

$$\Delta_b = \frac{3.5\sigma}{N^{1/3}}, \tag{4.78}$$

where σ is the sample standard deviation, and N is the sample size. This rule asymptotically minimizes the mean integrated square error (see eq. 4.14) and assumes that the underlying distribution is Gaussian; see [22]. An attempt to generalize this rule to non-Gaussian distributions is the Freedman–Diaconis rule,

$$\Delta_b = \frac{2(q_{75} - q_{25})}{N^{1/3}} = \frac{2.7\sigma_G}{N^{1/3}}, \tag{4.79}$$

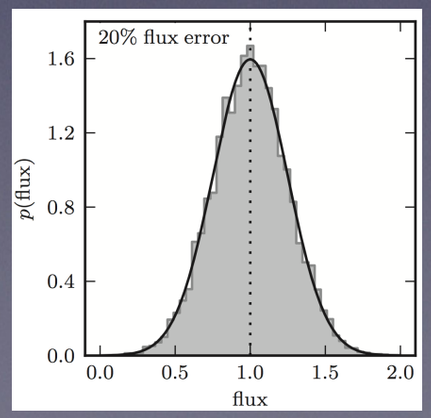
which estimates the scale ("spread") of the distribution from its interquartile range (see [12]). In the case of a Gaussian distribution, Scott's bin width is 30% larger than the Freedman–Diaconis bin width. Some rules use the extremes of observed values to estimate the scale of the distribution, which is clearly inferior to using the interquartile range when outliers are present.

Although the Freedman–Diaconis rule attempts to account for non-Gaussian distributions, it is too simple to distinguish, for example, multimodal and unimodal distributions that have the same σ_G

What is a histogram?

- ---> Data modeled by a step function
- —> How do we determine/estimate/guess the bin width?
- Do all the bins have to have the same width? No, there are methods to handle variable bin width (e.g. Bayesian Blocks Method, check out astroML mentioned below)
- Do we really have to bin data to estimate model parameters? No, we do not! (and should not). Binning should be used only for visualization.

A good example: to "fit" a Gaussian to data, simply estimate the location and scale parameters, as we just discussed. To show results, you can then use histogram and overplot the fit.



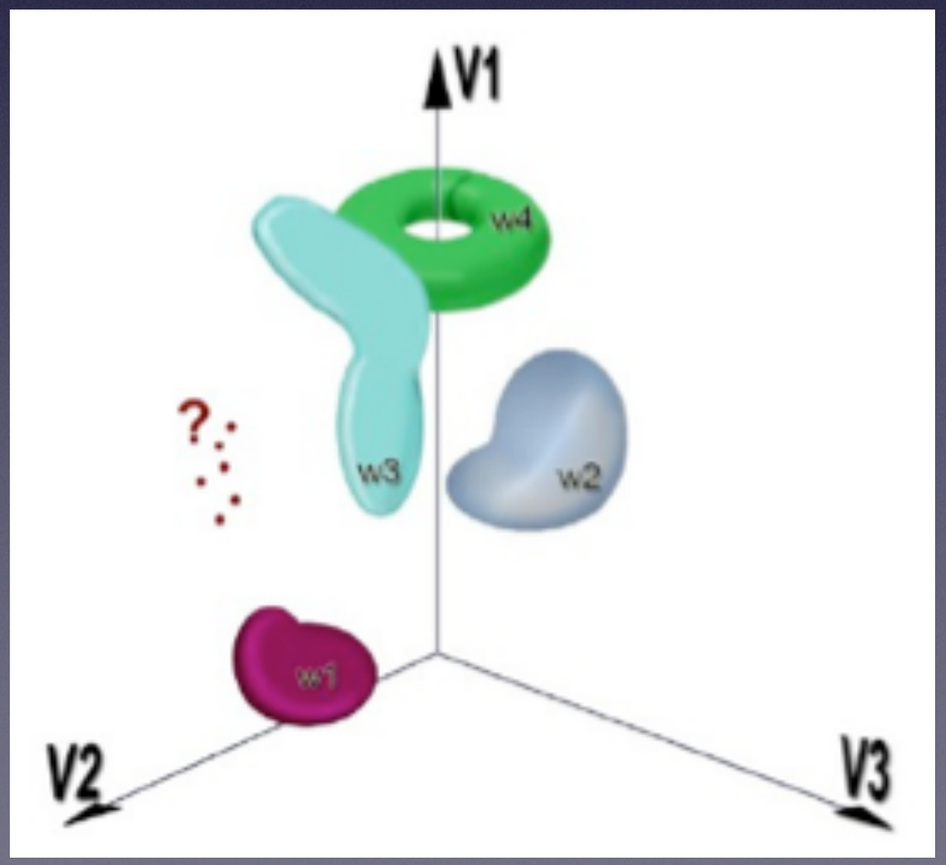
Outline

- **How to estimate location and scale?**
  - Mean, median, std. dev. and their errors
  - Central Limit Theorem
  - Robust statistics
  - Gaussianity, Chi-squared, and non-gaussianity
- **How to make a histogram?**
- **If we had more time…**

Statistical analysis of a massive LSST dataset

- A large (100 PB) database and sophisticated analysis tools: for each of 40 billion objects there will be about 1000 measurements (each with a few dozen measured parameters)

Data mining and knowledge discovery



- 10,000-D space with 40 billion points
- Characterization of known objects
- Classification of new populations
- Discoveries of unusual objects

Clustering, classification, outliers

Data analysis challenges in the era of Big Data

1) Large data volume

2) Large number of objects

3) Highly multi-dimensional space

4) Unknown statistical distributions

5) Time-series data

6) Truncated, censored and missing data

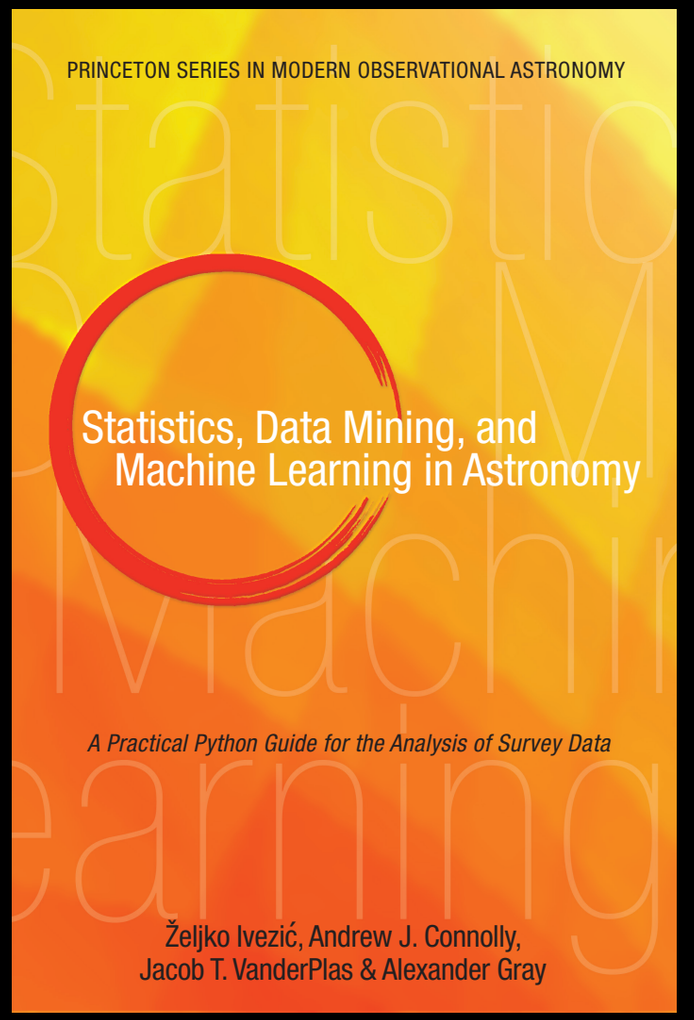
7) Unreliable quantities (e.g. unknown systematics and random errors)

How can efficiently train astronomy/physics students to use sophisticated methods from statistics, data mining and machine learning?

Statistics, Data Mining and Machine Learning in Astronomy

Željko Ivezić, Andrew Connolly, Jacob Vanderplas, Alex Gray

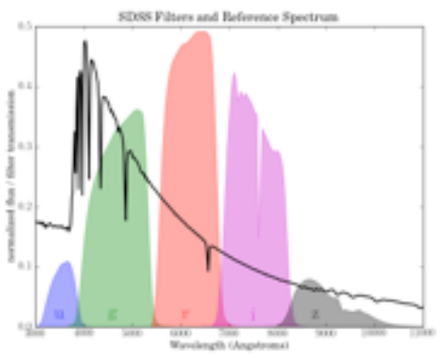
Princeton University Press, 2013

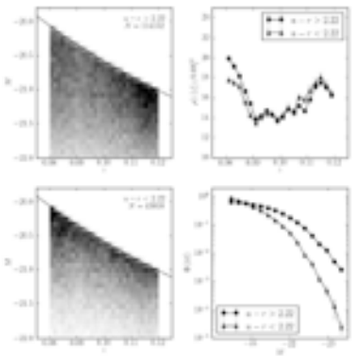


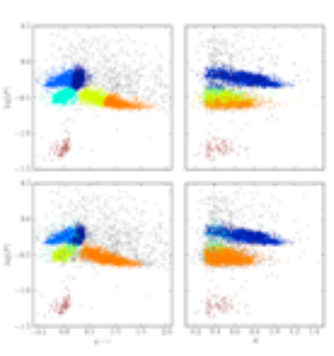
- Complete *Practical* guide to statistical analysis, data exploration, and machine learning
- Example-driven approach, using real data (SDSS, LIGO, LINEAR, WMAP, and others)
- All book figures and examples generated in python (matplotlib), with code available online – for free!
- Makes use of *numpy*, *scipy*, *matplotlib*, *scikit-learn*, *pymc*, *healpy*, and others
- Supporting python package: *astroML*

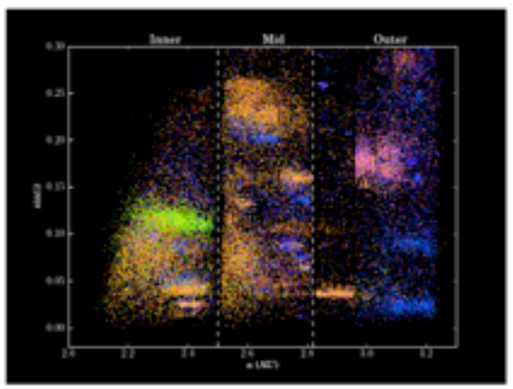
New book

AstroML: Machine Learning and Data Mining for Astronomy









News

October 2012: astroML 0.1 has been released! Get the source on Github

Our Introduction to astroML paper received the CIDU 2012 best paper award.

Links

astroML Mailing List

GitHub Issue Tracker

Videos

Scipy 2012 (15 minute talk)

Citing

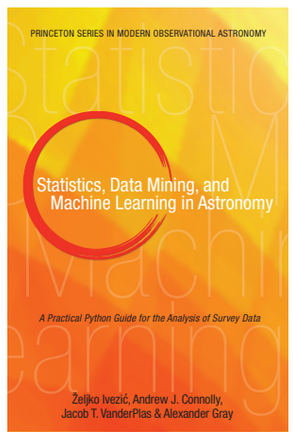
If you use the software, please consider citing astroML.

AstroML is a Python module for machine learning and data mining built on numpy, scipy, scikit-learn, and matplotlib, and distributed under the 3-clause BSD license. It contains a growing library of statistical and machine learning routines for analyzing astronomical data in python, loaders for several open astronomical datasets, and a large suite of examples of analyzing and visualizing astronomical datasets.

Downloads

- Released Versions: Python Package Index
- Bleeding-edge Source: github

The goal of astroML is to provide a community repository for fast Python implementations of common tools and routines used for statistical data analysis in astronomy and astrophysics, to provide a uniform and easy-to-use interface to freely available astronomical datasets. We hope this package will be useful to researchers and students of astronomy. The astroML project was started in 2012 to accompany the book **Statistics, Data Mining, and Machine Learning in Astronomy** by Zeljko Ivezic, Andrew Connolly, Jacob VanderPlas, and Alex Gray, to be published in late 2013. The table of contents is available here: **`here(pdf)`**.



User Guide

1. Introduction

- 1.1. Philosophy

Open source!
www.astroML.org

• If we had more time…

- Correlation coefficients
- The bootstrap and the jackknife methods
- Maximum Likelihood Method
- The goodness of fit and model selection
- Bayesian statistics
- Markov Chain Monte Carlo methods
- Regression (“fitting”, LSQ, outliers)
- Density estimation (“multi-dimensional histograms”)
- Clustering
- Classification
- Dimensionality Reduction (PCA and friends)
- Time-series analysis (periodogram, stochastic processes)

• What is Bayesian statistics?

"As an undergraduate, I always found the subject of statistics to be rather mysterious. I was already familiar with the binomial, Poisson and normal distributions. Most of this made sense, but only seemed to relate to things like rolling dice, flipping coins, shuffling cards and so on. However, having aspirations of becoming a scientist, what I really wanted to know was how to analyse experimental data. Thus, I eagerly looked forward to the lectures on statistics.

Sadly, they were a great disappointment. Although many of the tests and procedures expounded were intuitively reasonable, there was something deeply unsatisfactory about the whole affair; there didn't seem to be any underlying principles! Hence, the course on 'probability and statistics' had led to an unfortunate dichotomy: probability made sense, but was just a game; statistics was important, but it was a bewildering collection of tests with little obvious rhyme or reason."

I felt exactly like Sivia!

- D.S. Sivia
(Data Analysis: A Bayesian Tutorial)

Bayes' Theorem:

$$p(M|D) = \frac{p(D|M)\,p(M)}{p(D)}$$

M = model
D = data

- **Bayesian statistics**

Posterior pdf for model M and parameters θ, given data D and prior information I

Bayes' Theorem:

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

The likelihood of data given M, with some fixed parameters θ, and I.

Probability of data ("normalization")

Prior

Probability for θ, given that M is true

Prior for model M

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

- **Bayesian statistics**

Bayes' Theorem:

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

This methodology can be used to derive all the results that we discussed in this lecture.
And much much more....