Data Mining In Modern Astronomy Sky Surveys:

Statistical Distributions in Astronomy

Ching-Wa Yip

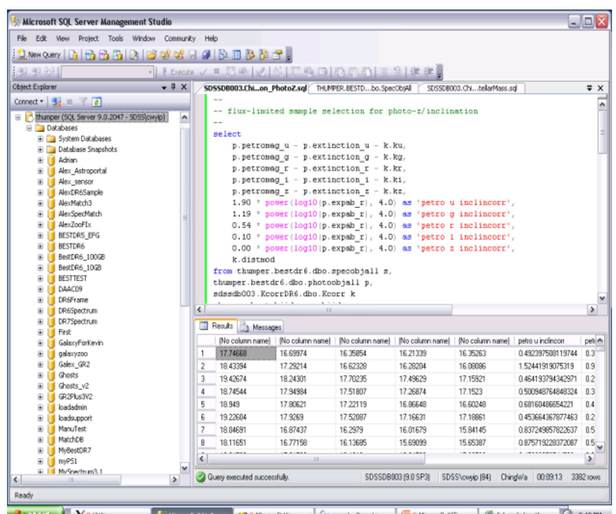
cwyip@pha.jhu.edu; Bloomberg 518

From Data to Information

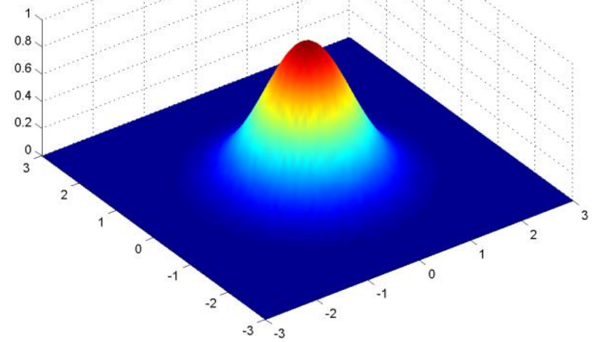
- We don’t just want data.
- We want information from the data.

Information ← **Database** ← **Sensors**

Data Analysis
or
Data Mining



Why Do We Use Statistics to Analyze Data?



- Describe the data:
  - What is the mean, median, mode?
  - What is the standard deviation?
  - What is the distribution?
  - What are the outliers?

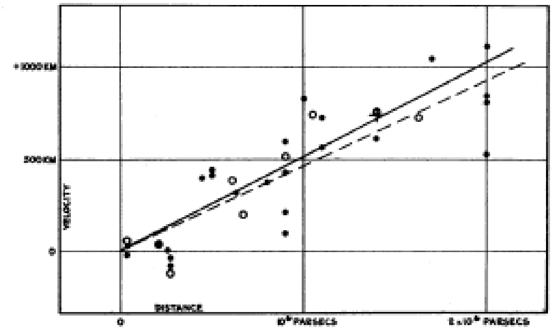
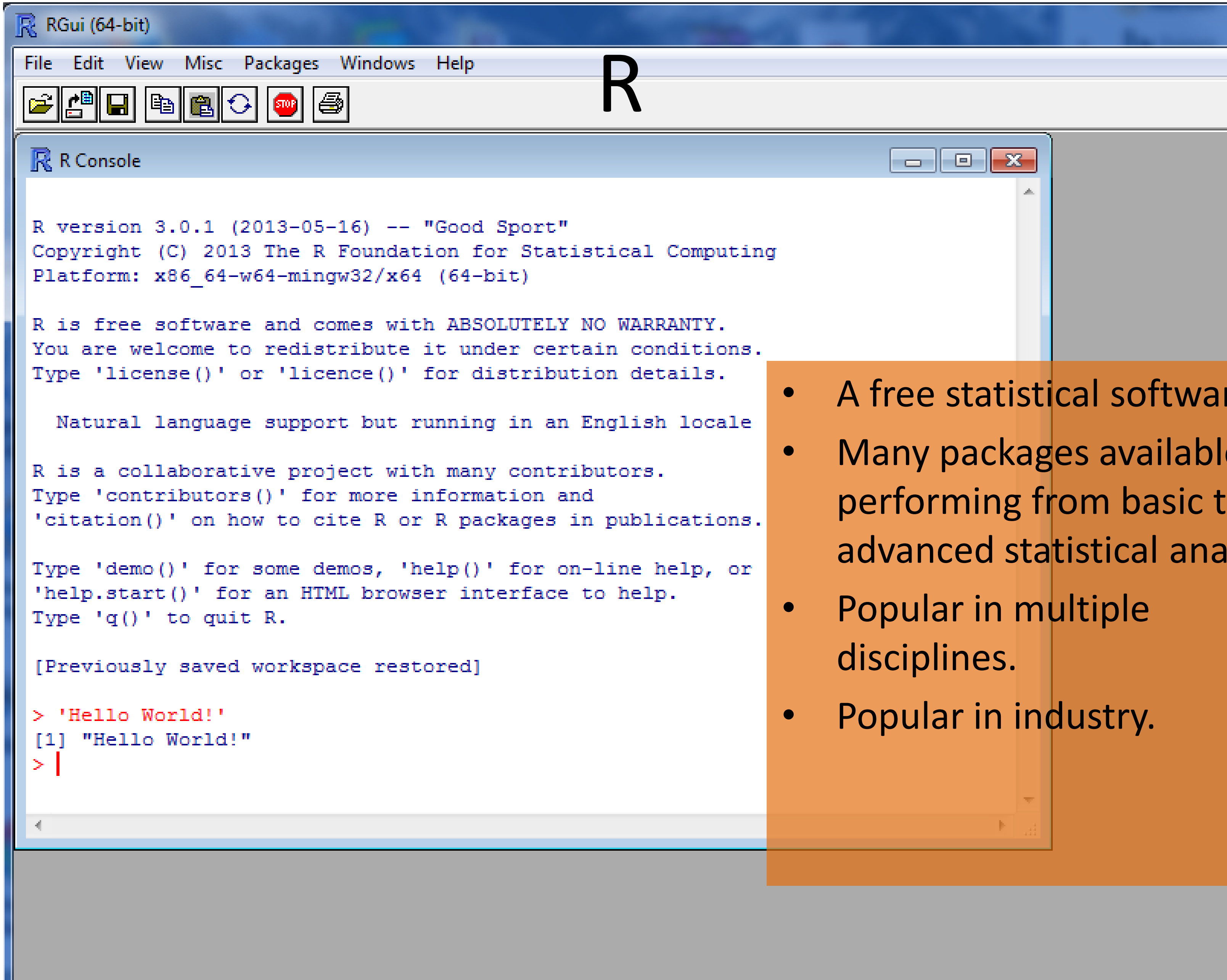


FIGURE 1
Velocity-Distance Relation among Extra-Galactic Nebulae.

- Find trends in the data:
  - Is there a correlation between two variables?
  - How well are they correlated?
  - What is the predicted value of a variable?

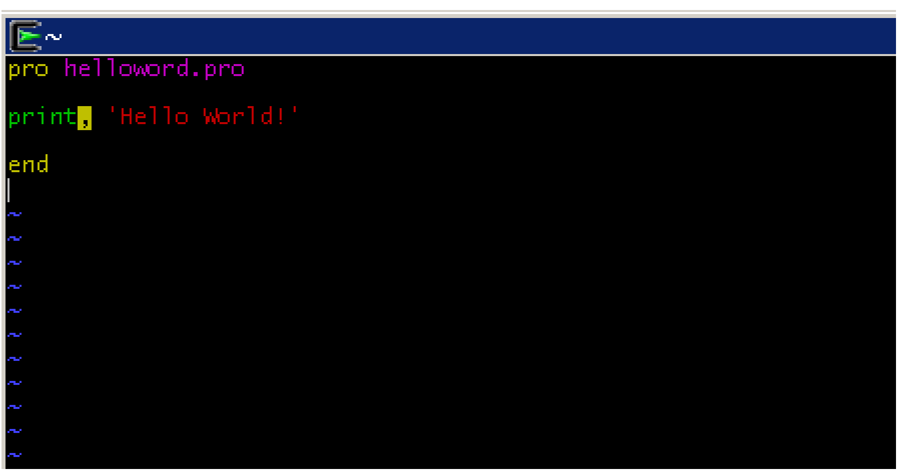
R

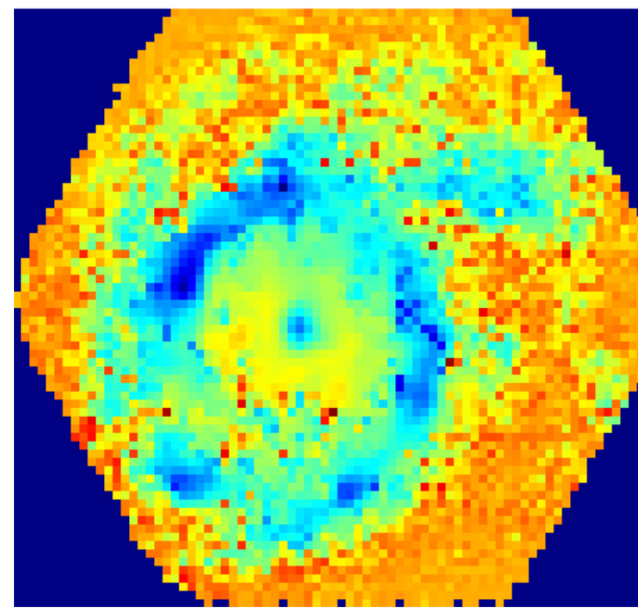


- A free statistical software.
- Many packages available for performing from basic to advanced statistical analyses.
- Popular in multiple disciplines.
- Popular in industry.

Interactive Data Language (IDL)

- A programming language for data analysis and plotting.
- Many *Procedures* for manipulating FITS file.
- Popular among astronomers.





(Age map of a galaxy.)

Other Programming Languages & Resources

- Python (free; getting popular among astronomers)
- Matlab
- C/C++/C#
- Java
- Numerical Recipes (William Press et al.)
  - Performs many numerical calculations
  - Can be used with different programming languages
- LAPACK
  - Performs algebraic and matrix calculations
  - Can be used with different programming languages

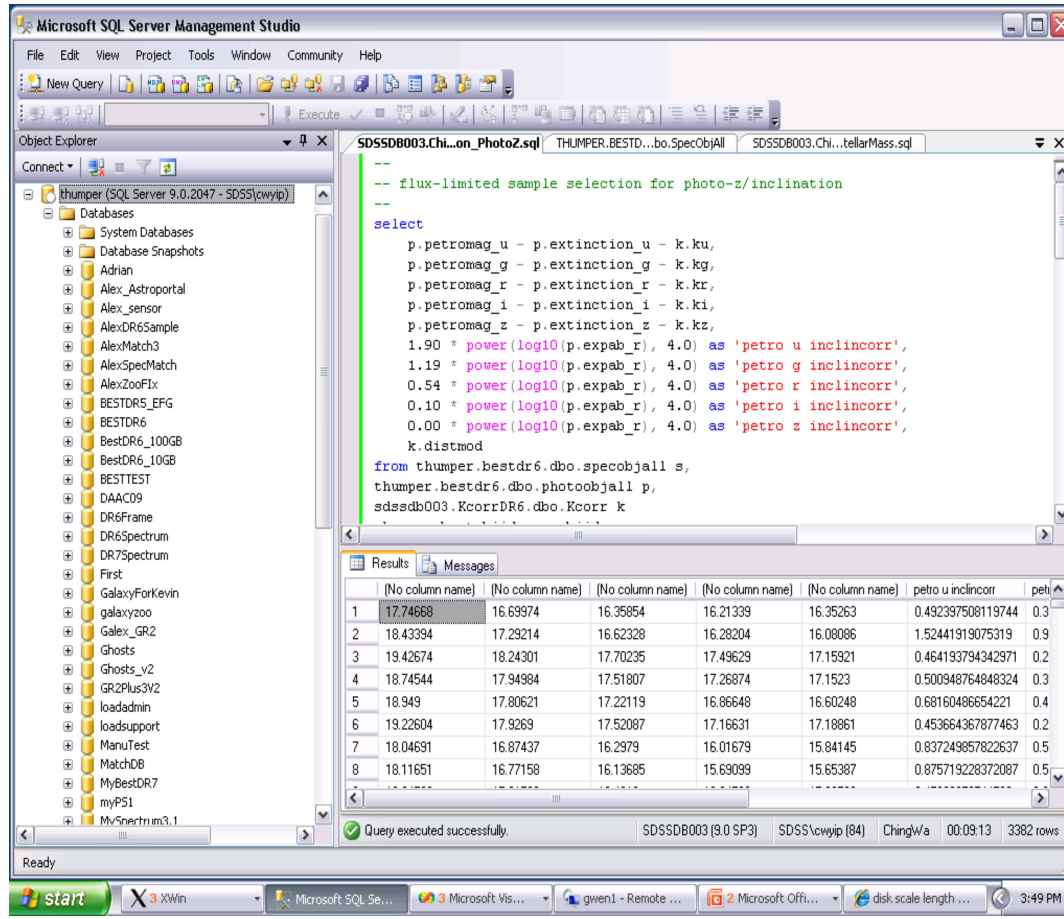
Future: Data Analysis using Database

- Automated data analysis:

Select data from DB using C# routines with SQL scripts embedded

Perform computations

Output results to DB, if necessary



(MS SQL Server. Source: Alex Szalay)

Basic Description of the Data: Mean, Median, Mode

- Mean = average value
- Median = middle value
- Mode = most common value

E.g. 10 sampled values =

3.4	4.8	8.4	9.6	2.3	9.6	5.6	9.6	4.8	2.2
(3)	(4)	(7)	(8)	(2)	(9)	(6)	(10)	(5)	(1)

n = 10

Mean = Sum of values / n = 6.03

Median = (4.8 + 5.6) / 2

Mode =

Basic Description of the Data: Mean, Median, Mode

- Mean = average value
- Median = middle value
- Mode = most common value

E.g. 10 sampled values =

3.4	4.8	8.4	9.6	2.3	9.6	5.6	9.6	4.8	2.2
(3)	(4)	(7)	(8)	(2)	(9)	(6)	(10)	(5)	(1)

n = 10

Mean = Sum of values / n = 6.03

Median = (4.8 + 5.6) / 2

Mode = 9.6

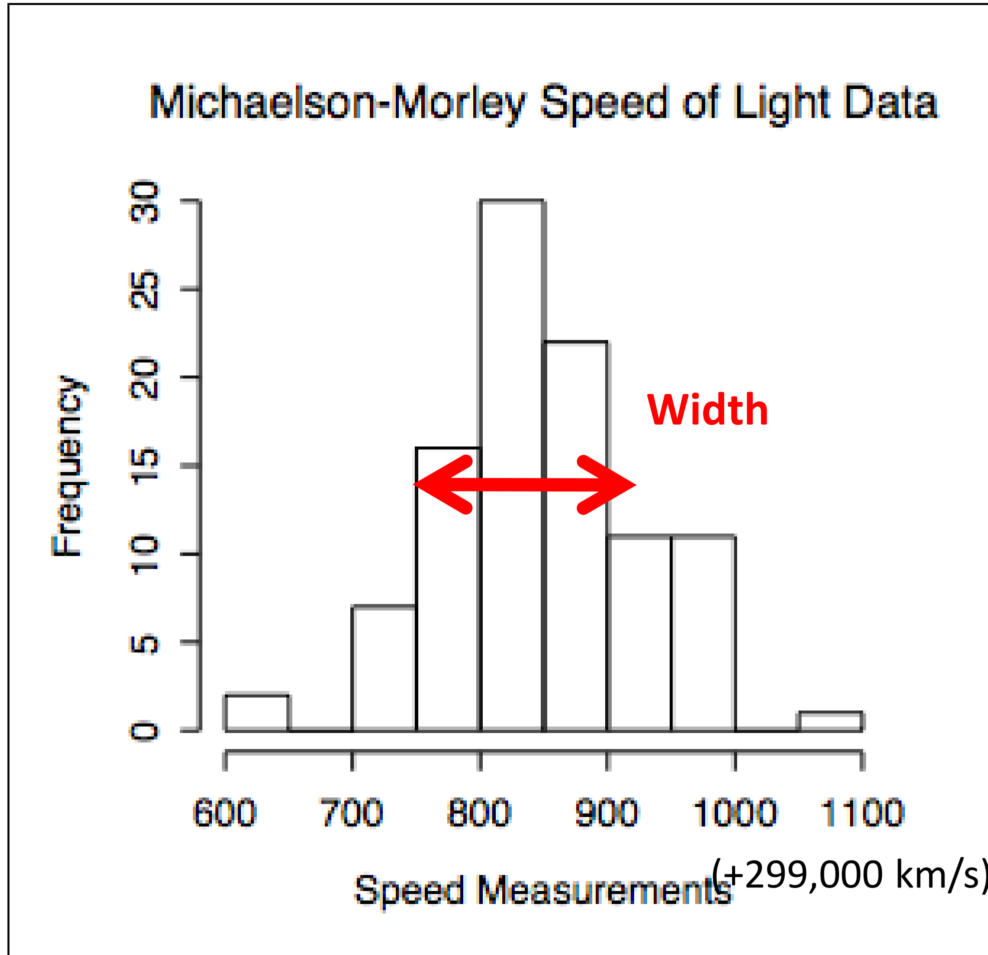
Width of a Distribution: Standard Deviation (or "SD", "σ")

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

n = The number of data points

$\bar{x}$ = The mean of the x_i

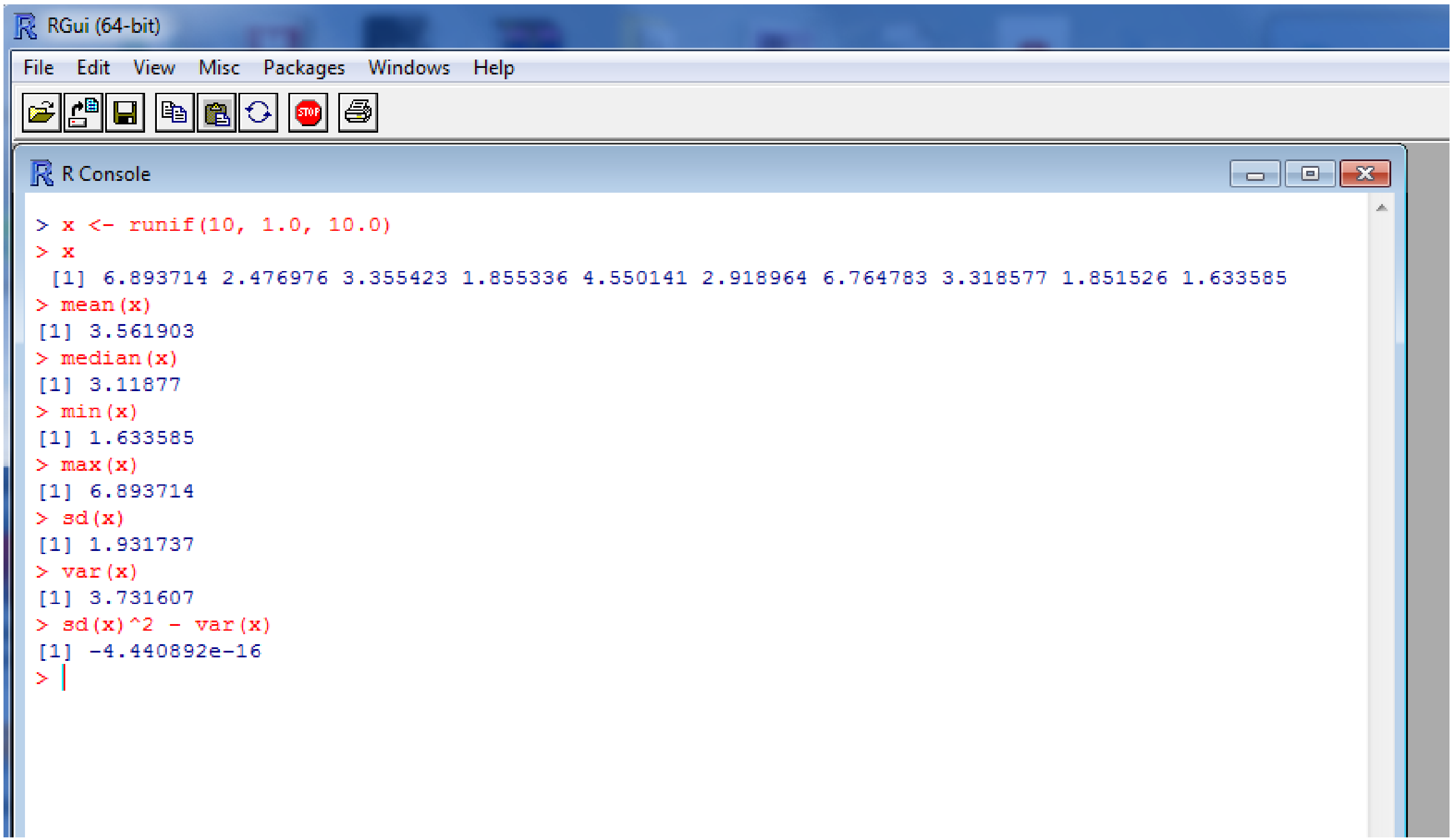
x_i = Each of the values of the data



Shape of a Distribution

- In general, sampled values may not form a well-quantified distribution (or, not analytical).
- We can use the binned histogram to report the shape of the distribution.
- Advanced: We can fit a linear combination of multiple functions.

Basic Statistics in R



Key Concepts in Statistics

- Variables
- Population Distribution vs. Sampling Distribution
- Central Limit Theorem (CLT)

A Big Bag of Marbles

- Suppose we want to measure the average and variance of the weight of a big bag of marbles. How to do it?

We draw a *sample* from the *population*.

Random Variable (X)

- X's are *Independent*: the outcome of any one experiment does not influence the outcomes of others (random draws).
- X's are *Identically Distributed*: every X is drawn from the same distribution (same bag of marbles).

	X_1	X_2	X_3	X_4	…	X_{n-1}	X_n
E.g.,	1.1g	0.8g	1.0g	0.9g		0.7g	1.2g

Random Variable (X)

- X's are *Independent*: the outcome of any one experiment does not influence the outcomes of others (random draws).
- X's are *Identically Distributed*: every X is drawn from the same distribution (same bag of marbles).

	X_1	X_2	X_3	X_4	…	X_{n-1}	X_n
E.g.,	1.1g	0.8g	1.0g	0.9g		0.7g	1.2g

IID Variable

Types of Random Variables

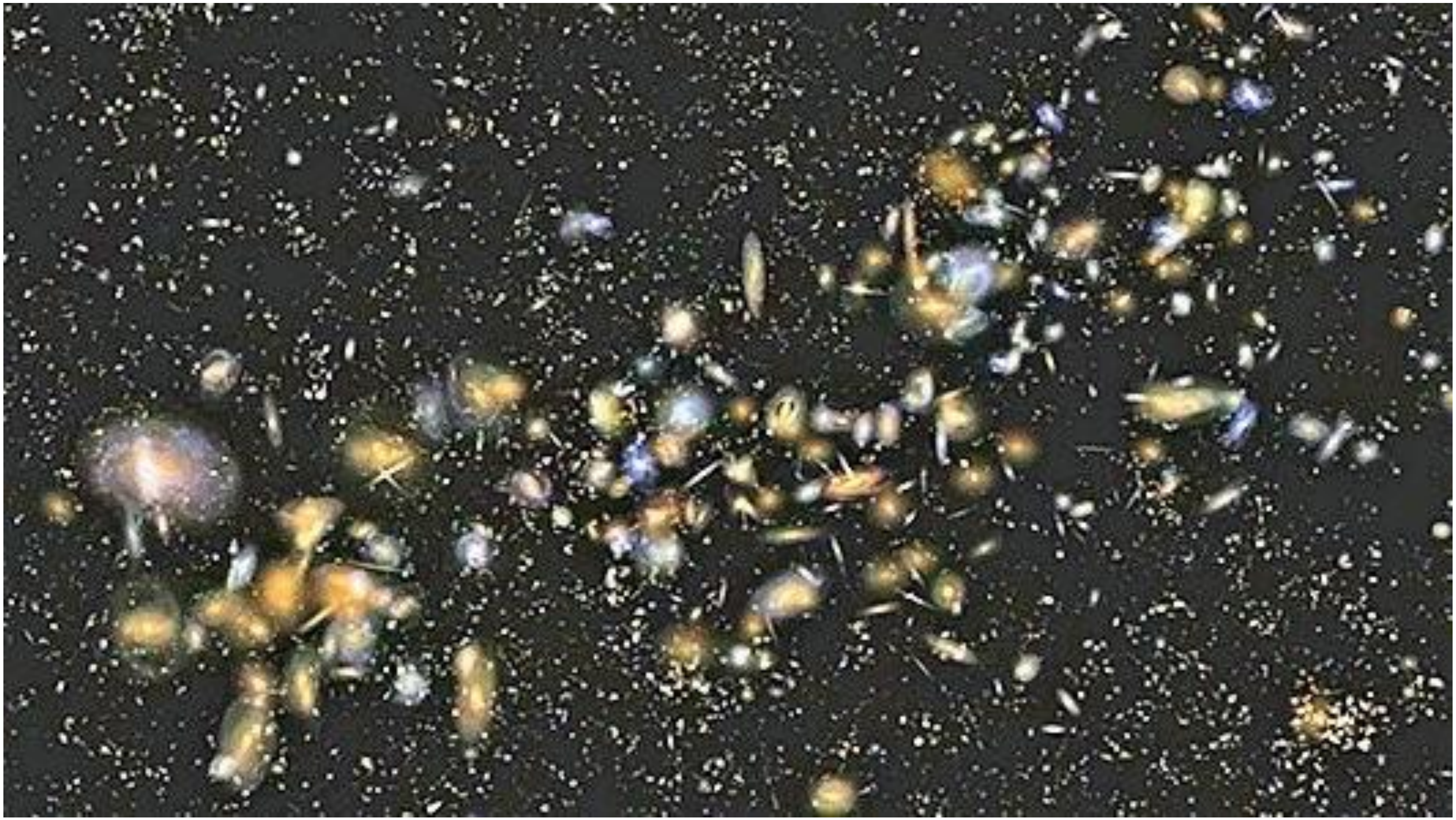
- Discrete variables
  - E.g., Photon Count: 10, 1, 6, 2, 14, 3, 5, 7, …
  - E.g., Binary States: 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, …
- Continuous variables
  - E.g., Luminosity (in units of solar luminosity):

    0.14, 2.46, 1.57, 4.52, …

Height of Trees in Forest

Sample
Vs.
Population

Color of Galaxies in Universe



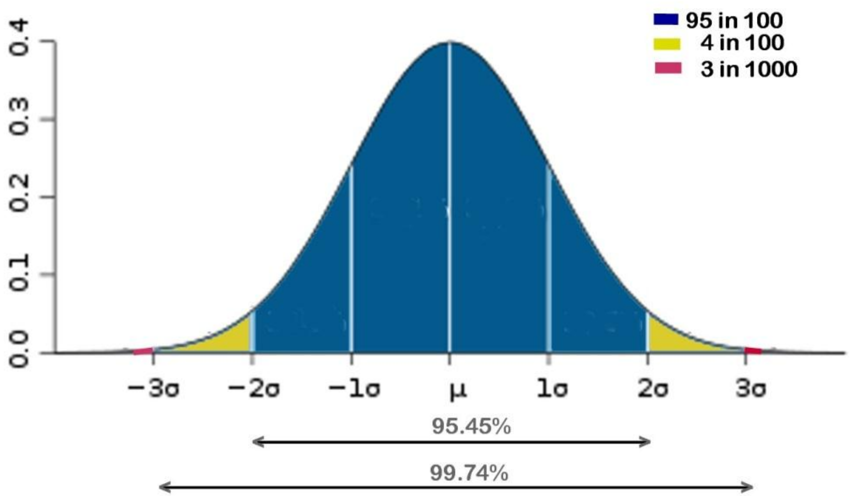
Credit: Miguel Angel Aragon Calvo

Statistical Distributions

- There are many statistical distributions available to describe experimental results.
- The most common ones in Astronomy are:
  - Gaussian Distribution
  - Poisson Distribution
  - Planck Distribution

Gaussian Distribution (or "Bell Curve", "Normal Distribution")

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

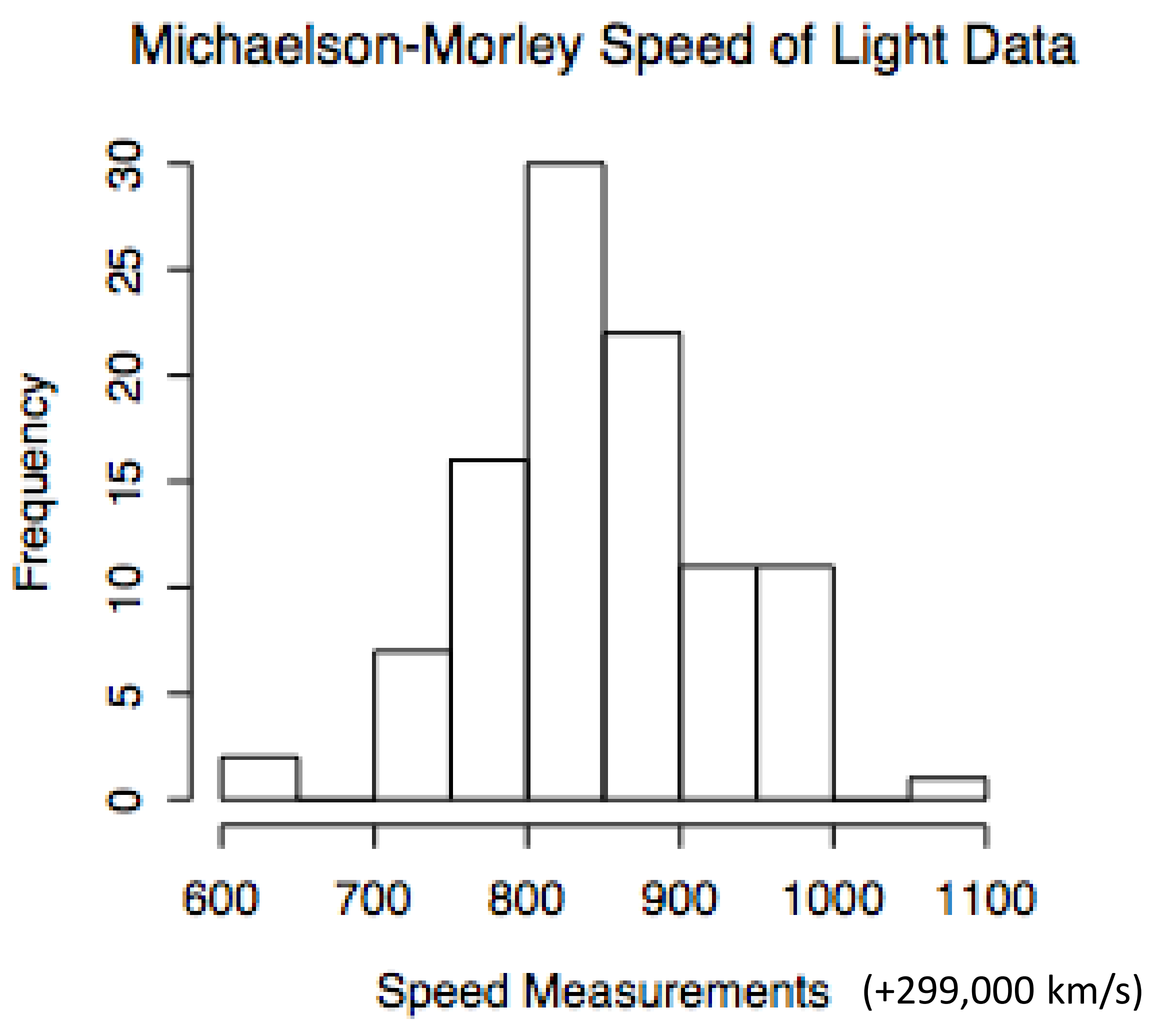


(Carl Friedrich Gauss, 1777-1855)

Uncertainty in Physical Measurements

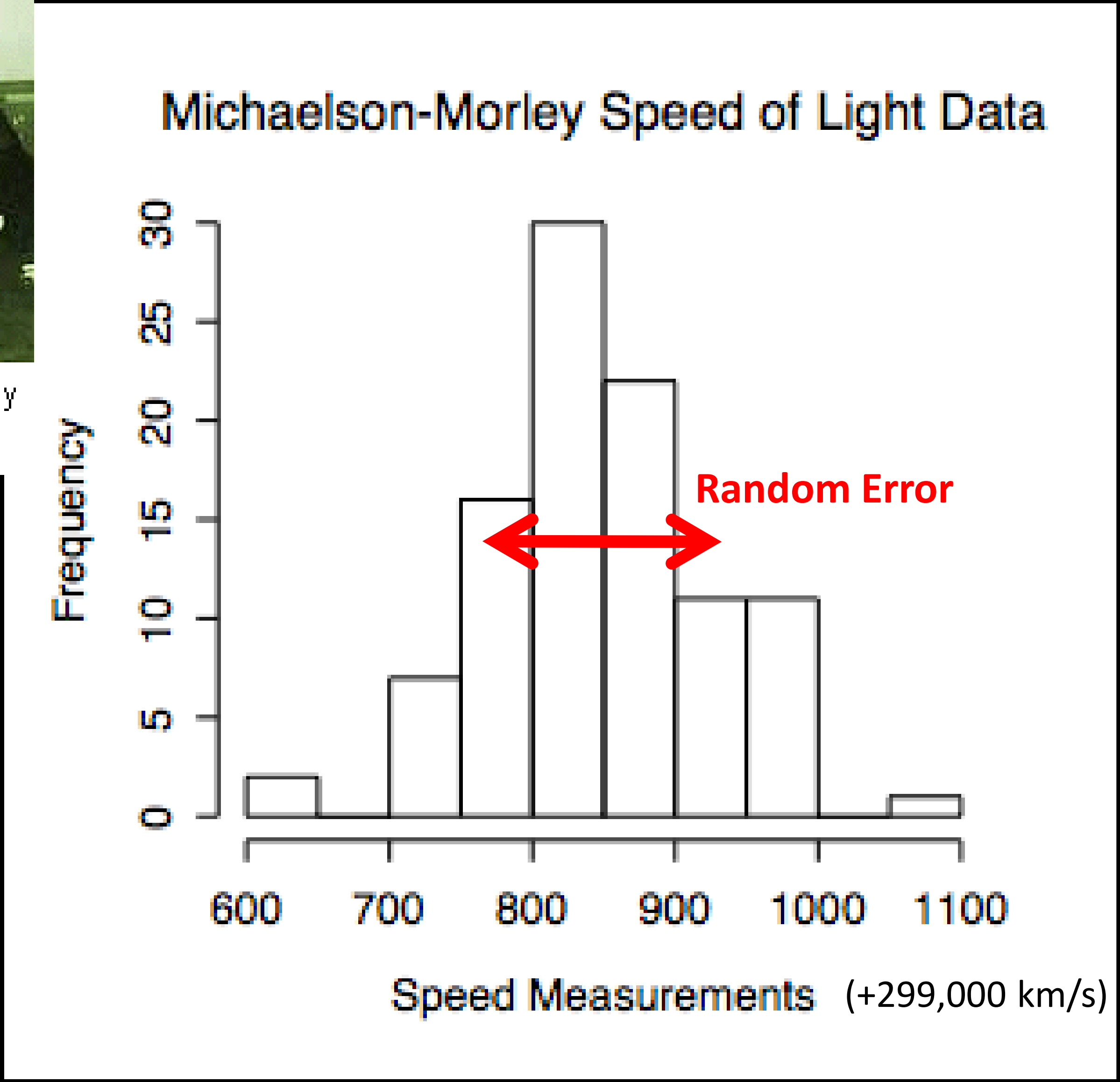
- For many experiments and observations concerning physical phenomena, we find that performing the procedure twice under (what seem!!!) identical conditions results in two different outcomes.
- *Such kind of outcomes follows a Gaussian Distribution.*
- For example: Michelson and Morley's speed-of-light experiment.

Albert Michelson at Chicago University
Photo: University of Chicago



(Credit: Wikipedia)

Albert Michelson at Chicago University
Photo: University of Chicago



(Credit: Wikipedia)

Main Lesson

- Michelson and Morley repeated the experiment many times to estimate the speed-of-light.
- They only knew the mean and SD of the population (X_i) exist, but they did not know of the values.
- It works because of the *Central Limit Theorem*.

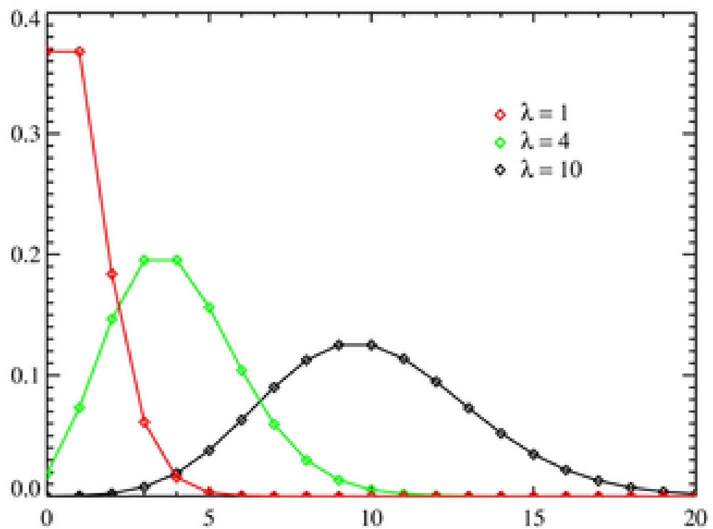
**Read first two paragraphs of the Handout.*

Poisson Distribution

- The probability of the number of events (occurring in a fixed period of time) given a known, expected count.

$$f(k;\lambda) = \frac{e^{-\lambda}\lambda^k}{k!},$$

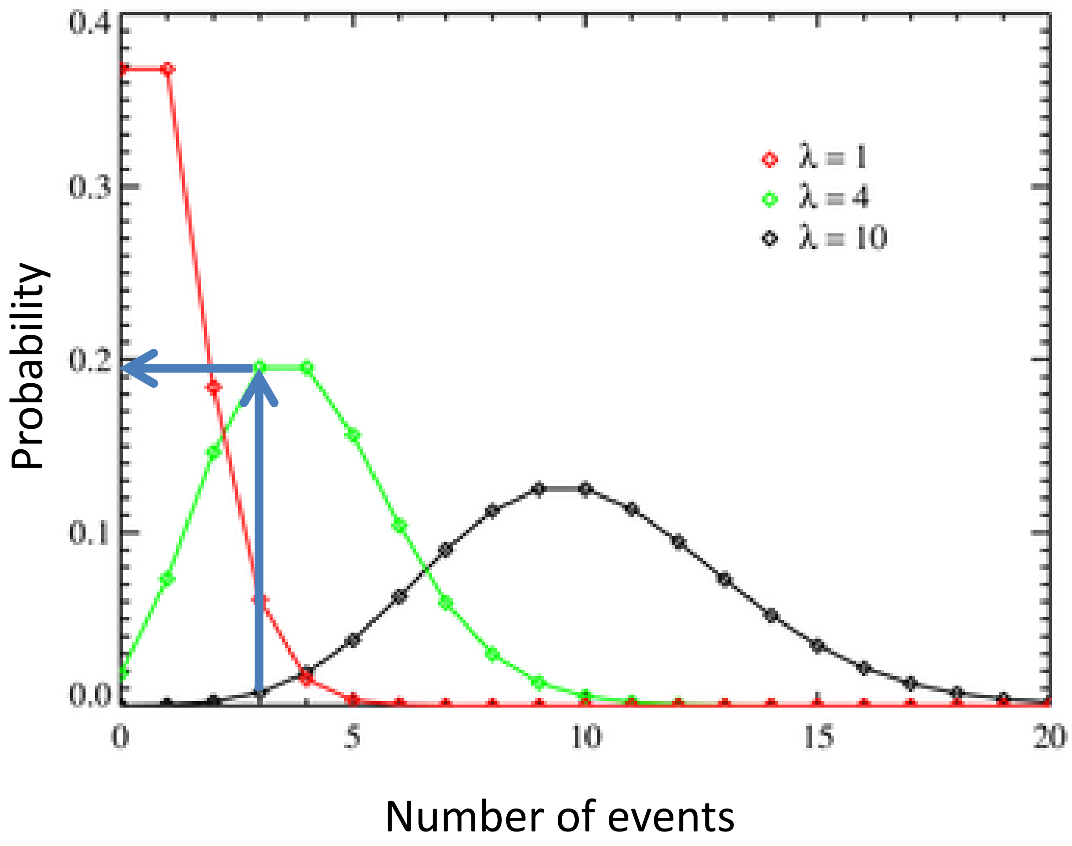
Probability



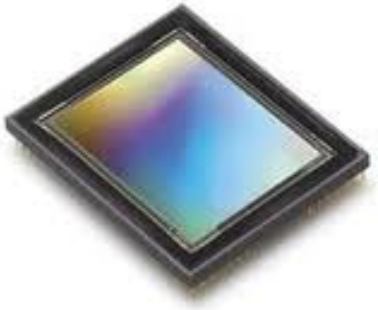
k (= number of events)

(Siméon Poisson, 1781-1840)

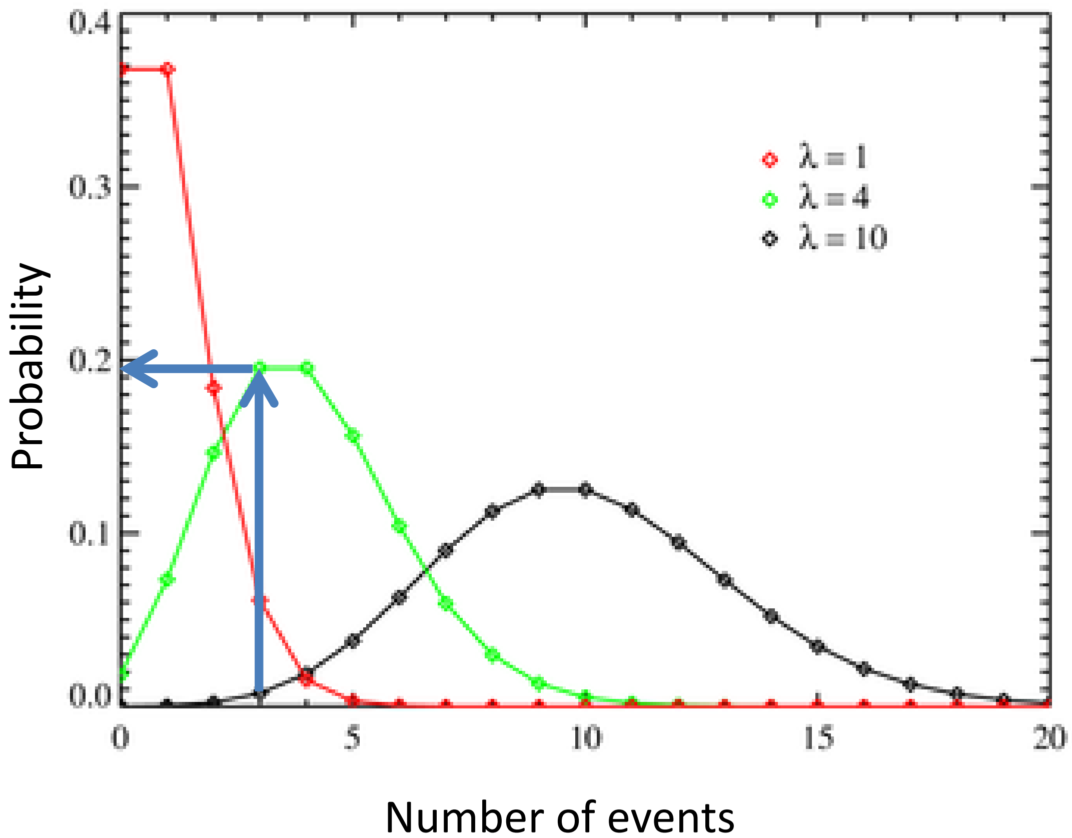
Examples of Poisson Distributions



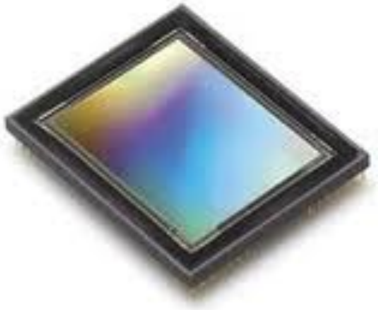
- I know from experience that I get 4 phone calls a day (λ = 4).
- Then there is about 20% chance I will get 3 phone calls today.



Examples of Poisson Distributions



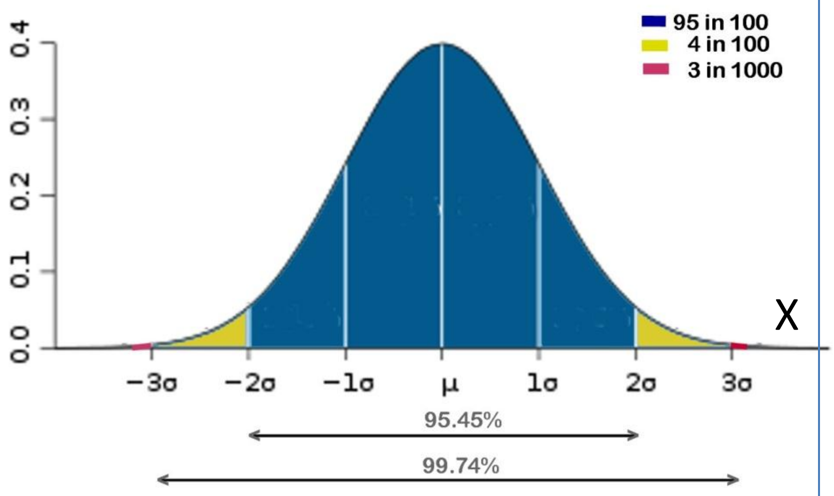
- I know from experience that I get 4 *photons* a day ($\lambda = 4$).
- Then there is about 20% chance I will get 3 *photons* today.



Probability: Chance of Occurrence

- Discrete Variable (e.g., 1, 2, 3, 4, 5, 6)

- Continuous Variable
  - E.g., Experimental results follow a Gaussian (also called Normal)

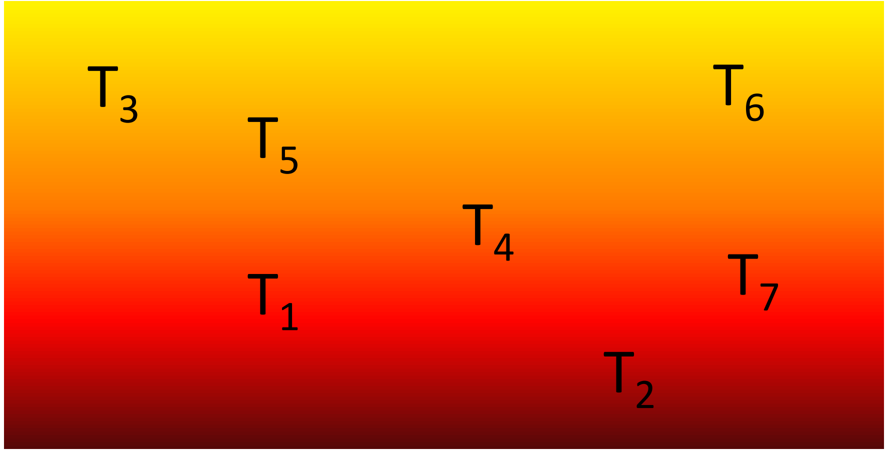


Probability of getting X between two values is the area under the Standard Normal Distribution between the two said values.

A Standard Normal Distribution is a Normal Distribution with total area = 1.

Random Number: Gaussian Distribution as a Case Study

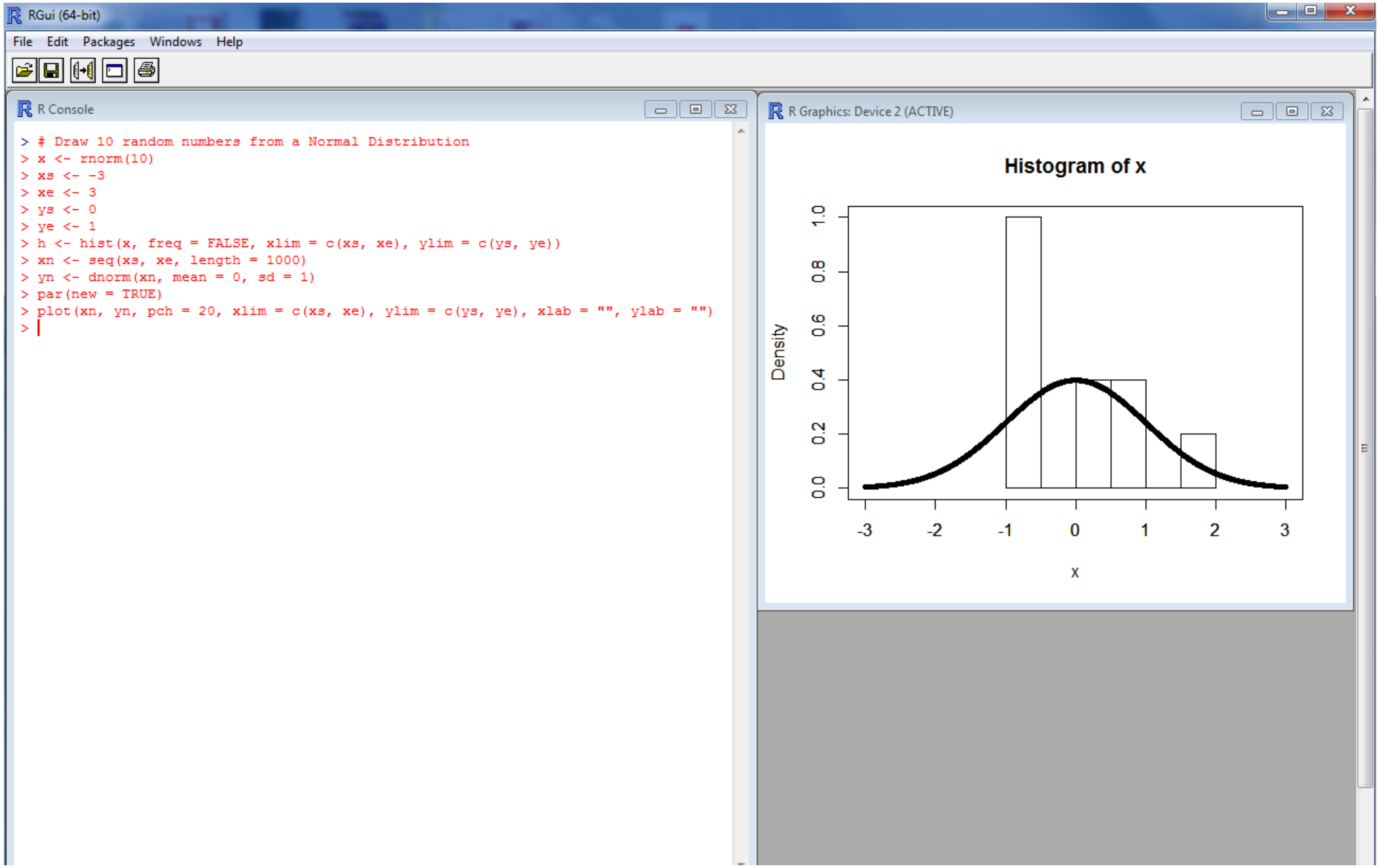
- When we analyze big datasets, *sampling methods* are commonly used.
- Random number plays an important role in sampling strategies.



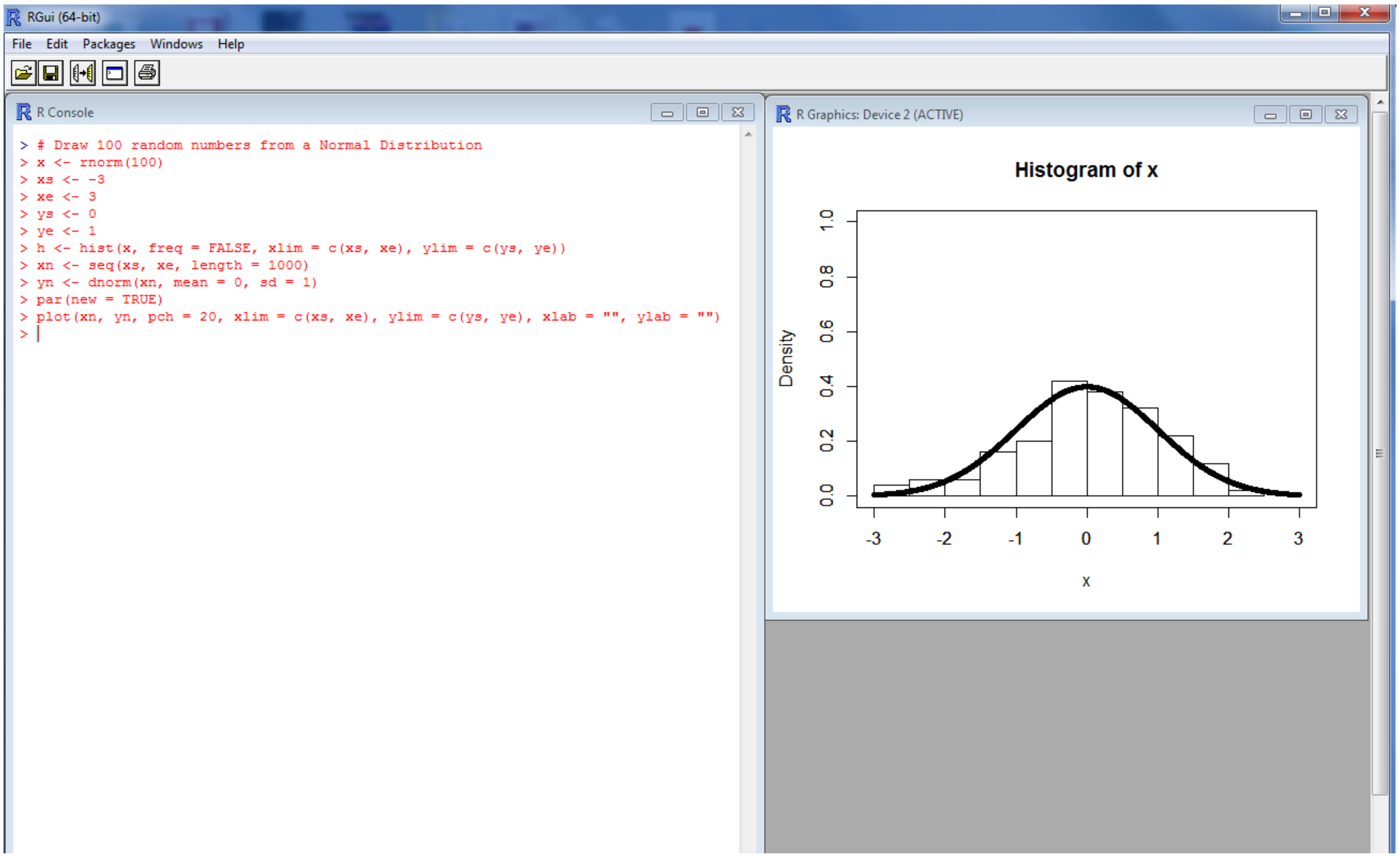
Suppose we have a piece of metal with temperature measurement T(x, y).

We want to estimate the average temperature without looking at all data (i.e., the population).

Distribution of 10 Random Numbers from Normal Distribution



Distribution of 100 Random Numbers from Normal Distribution

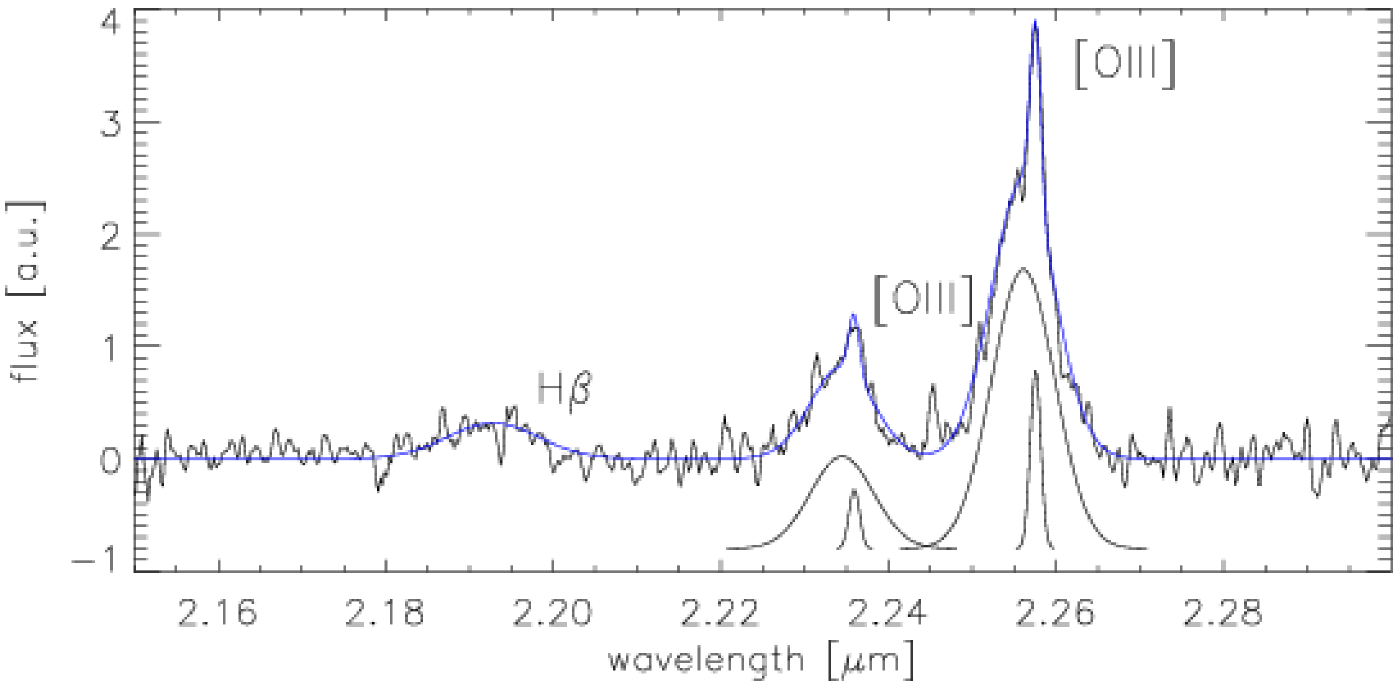


Distributions in Astronomy

- Gaussian Distribution
- Poisson Distribution
- Planck Distribution
- Etc.

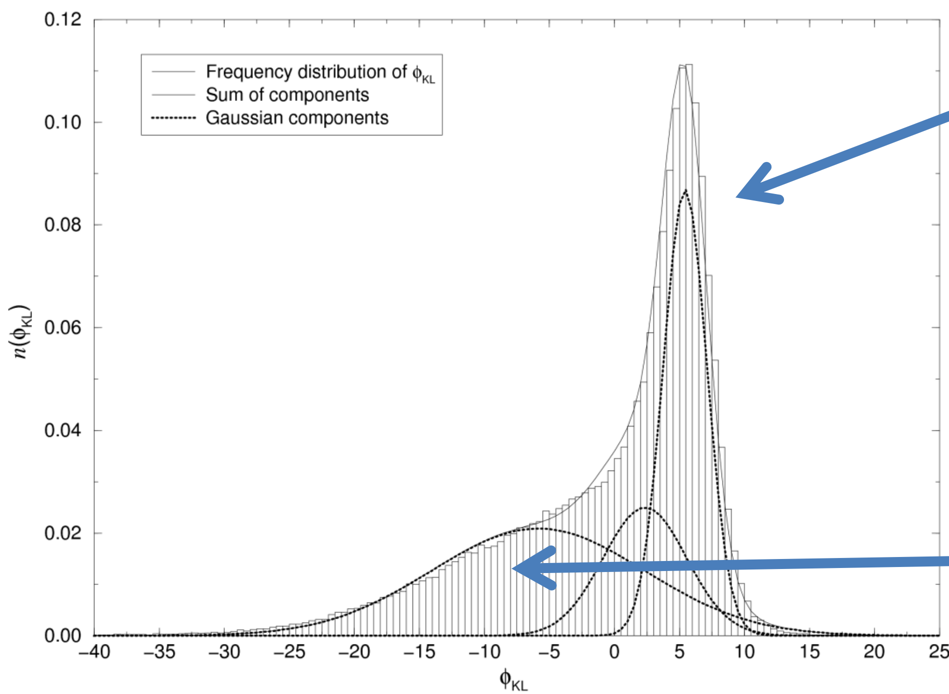
Emission Lines in Galaxies: Gaussian

- Fitting single or multiple Gaussian functions to the line profiles in a galaxy spectrum.



Colors of Galaxies in Nearby Universe: Multiple Gaussians

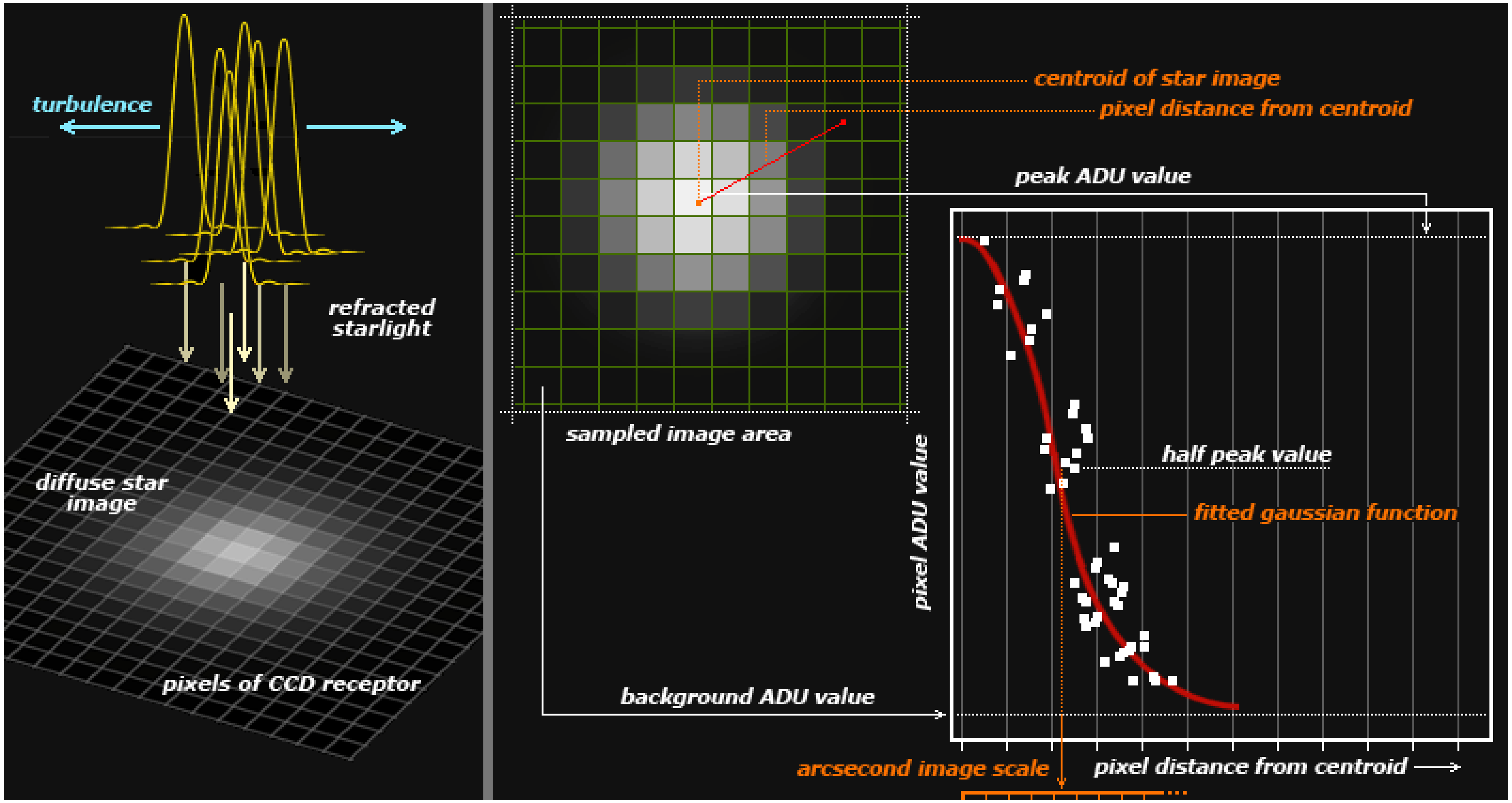
Frequency



Color (Redder →)

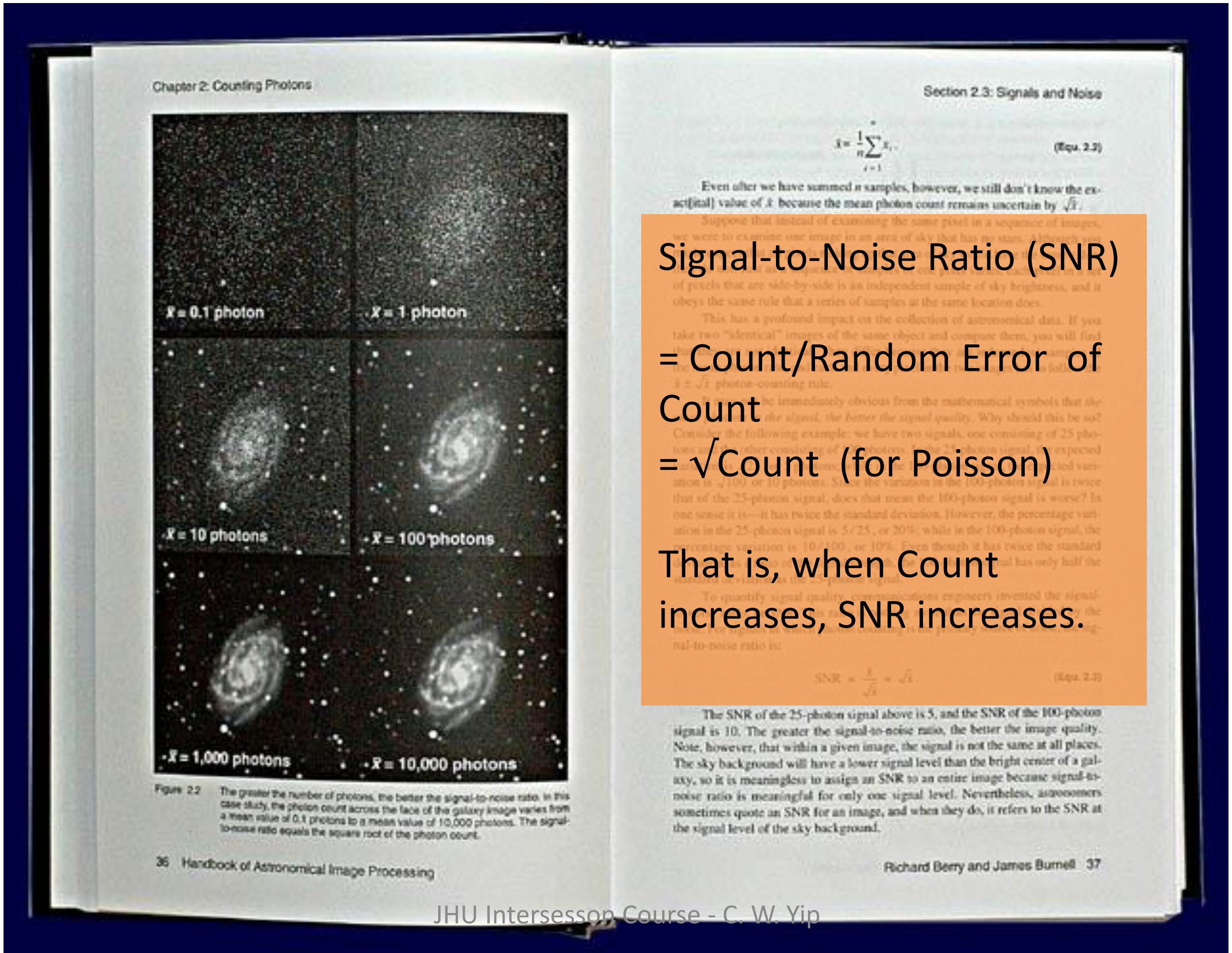
(Yip, Connolly, Szalay, et al. 2004)

Seeing Disk of Stars: Gaussian



Cause of Seeing Disk of Stars: Atmospheric Disturbance

Photon Counts in Astronomical Images: Poisson Distribution



Signal-to-Noise Ratio (SNR)

= Count/Random Error of Count

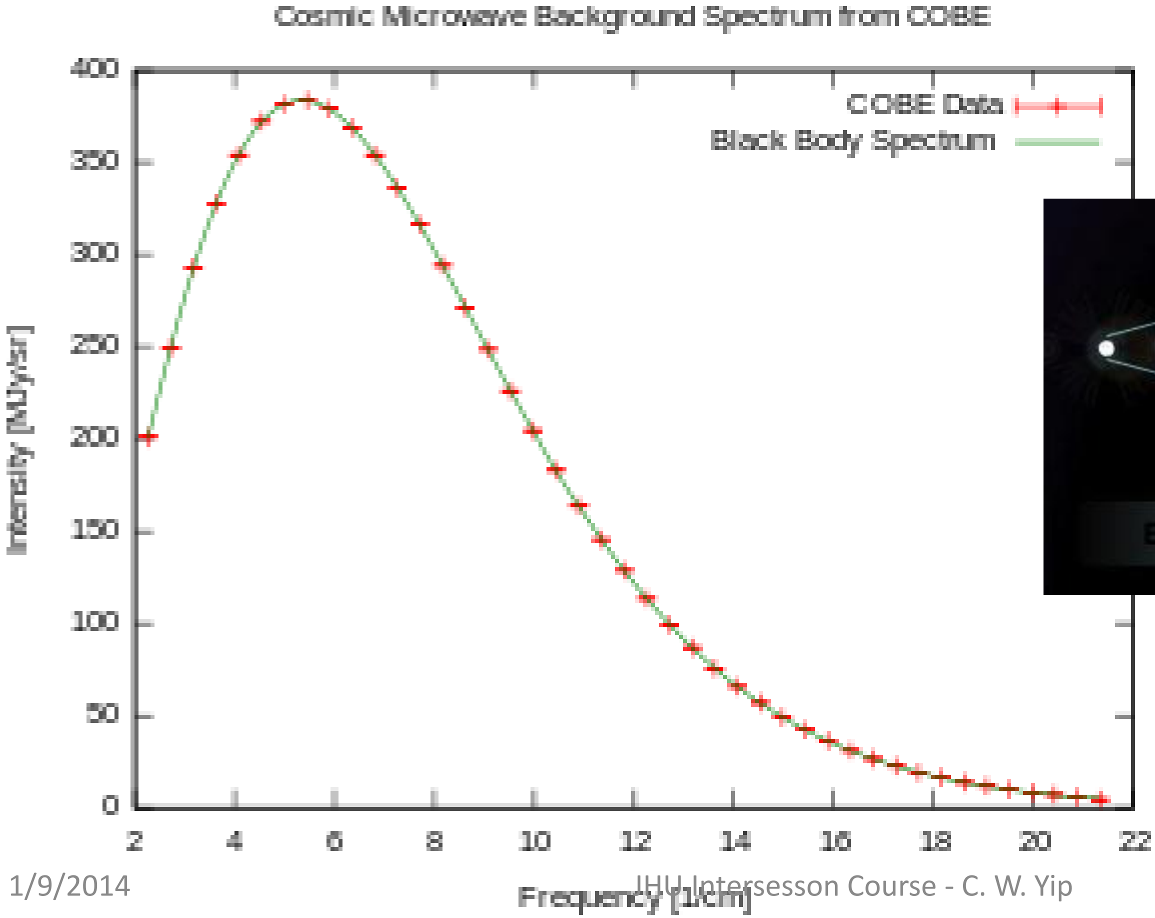
= √Count (for Poisson)

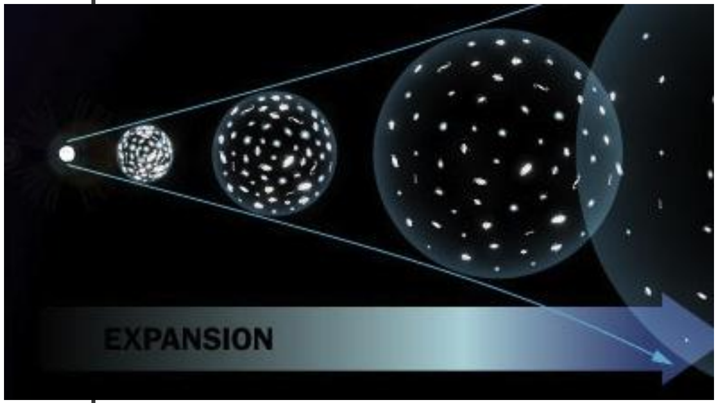
That is, when Count increases, SNR increases.

Cosmic Microwave Background:

Planck Distribution

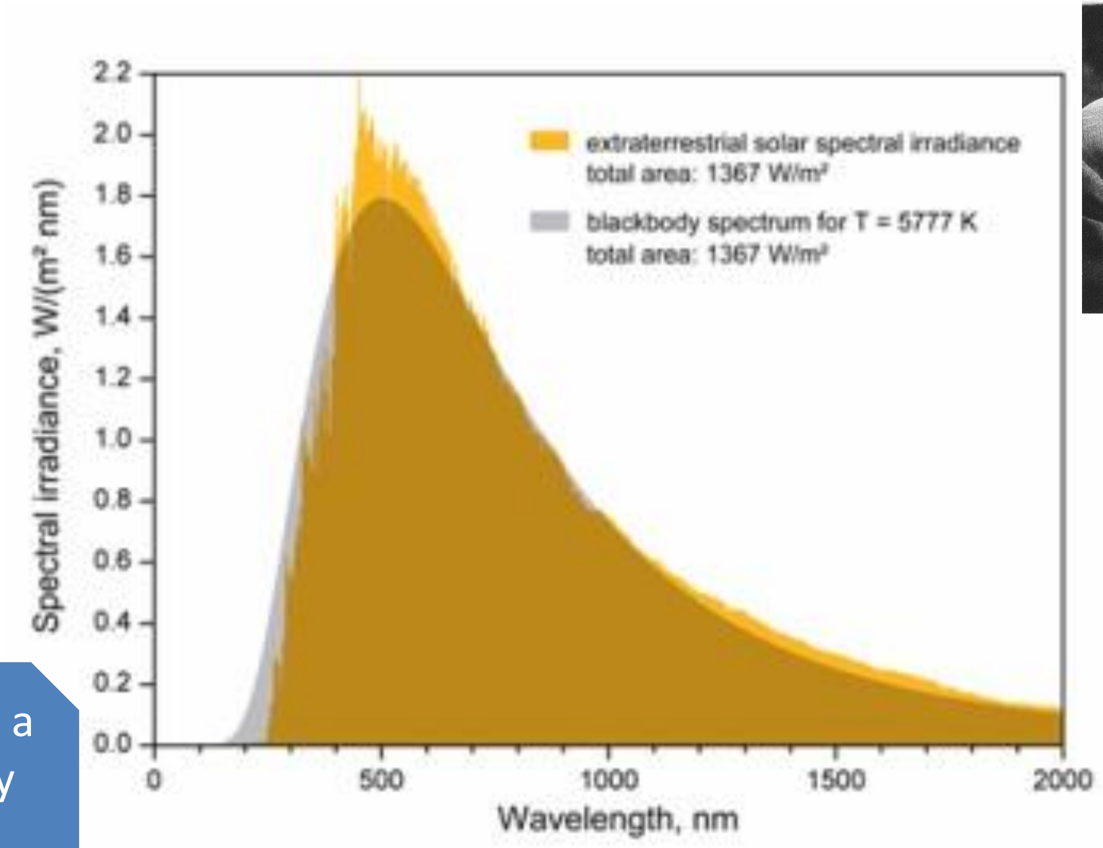
$$I(\nu) = \left(\frac{8\pi h\nu^3}{c^3}\right)\frac{1}{e^{h\nu/k_B T} - 1}$$

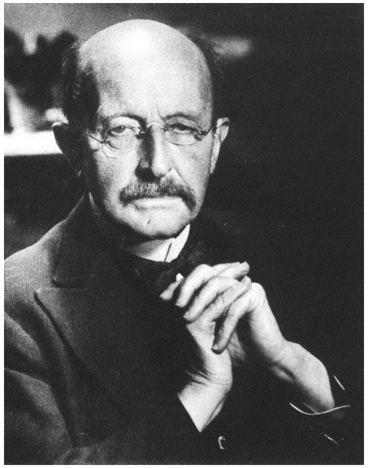




- Universe expands and cools down.
- The photons were hot, but cold now (3 Kelvin).

Black Body Radiation from the Sun: Planck Distribution





(Max Planck, 1858 – 1947)

Why is the Sun not a perfect Black Body (or Planck Distribution)?