

Data Mining In Modern Astronomy Sky Surveys:  
*Statistical Distributions in Astronomy*

Ching-Wa Yip

[cwyip@pha.jhu.edu](mailto:cwyip@pha.jhu.edu); **Bloomberg 518**

# From Data to Information

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

Information



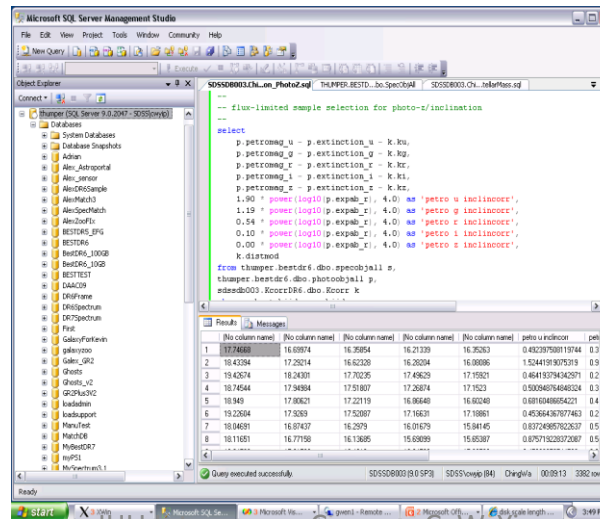
Database



Sensors

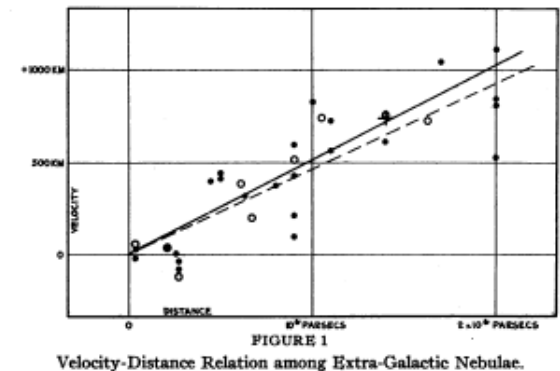
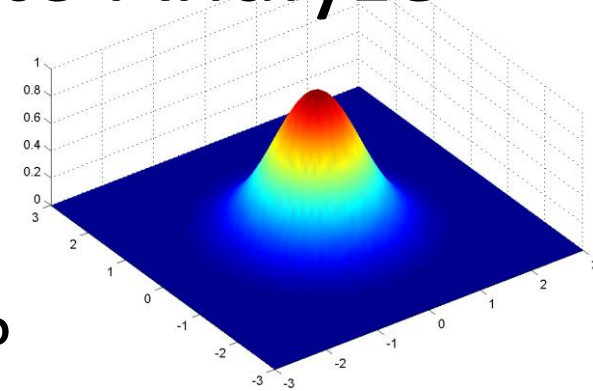


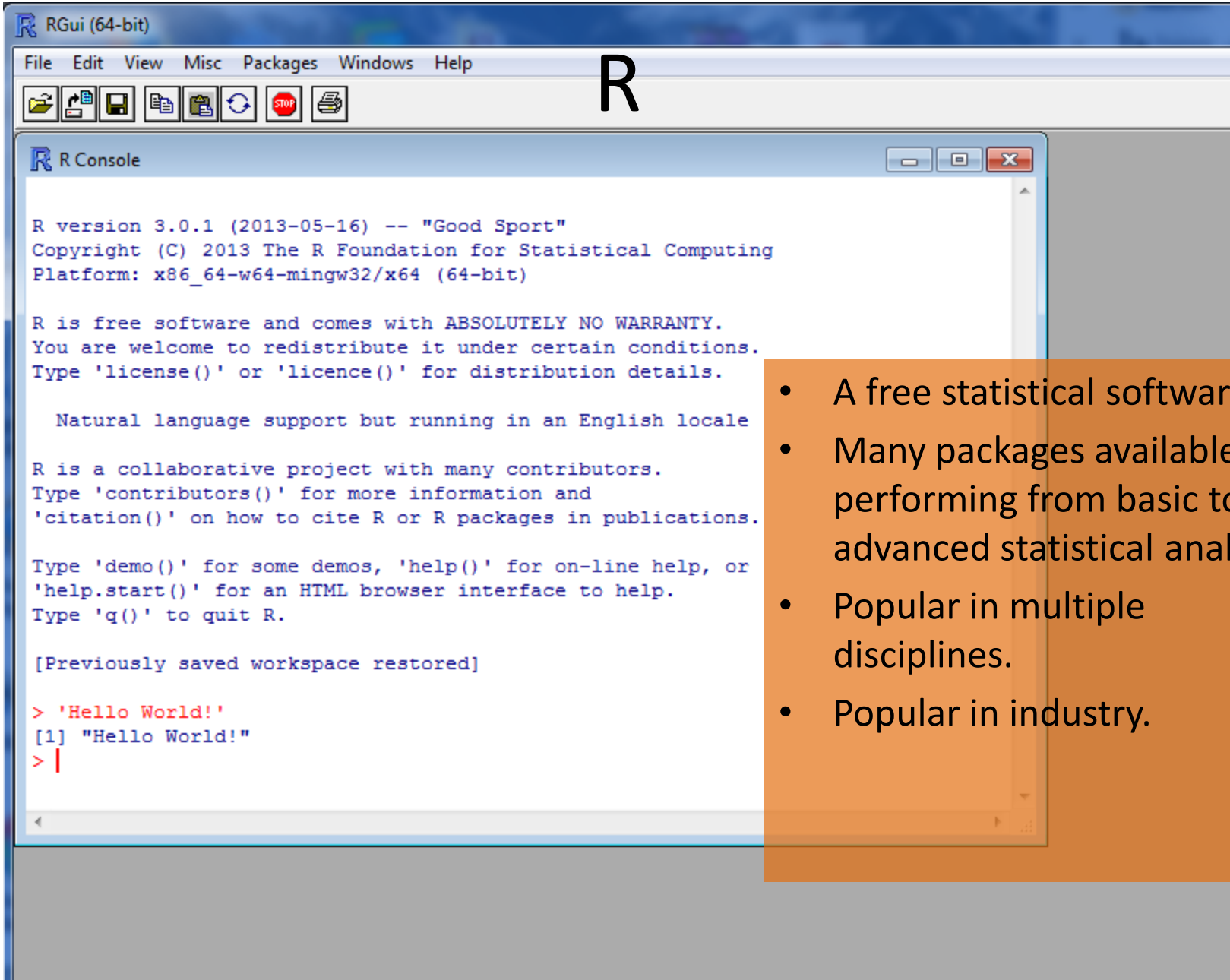
Data Analysis  
or  
Data Mining



# 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?
- 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?



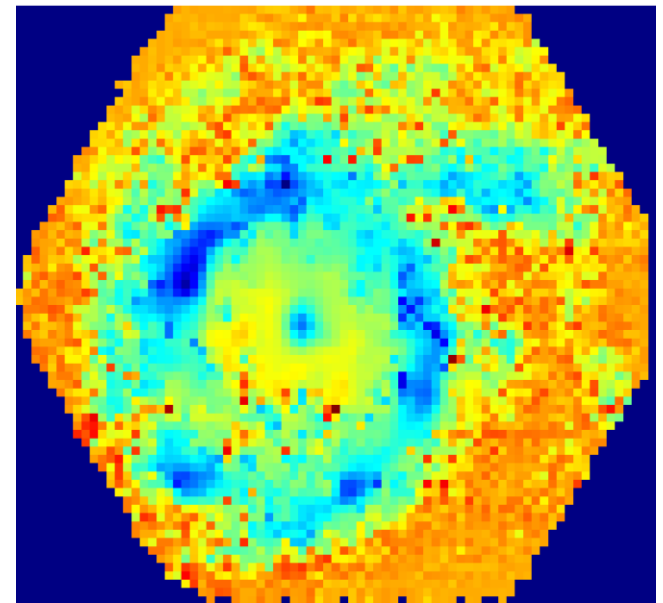


- A free statistical software.
- Many packages available for performing from basic to advanced statistical analyses.
- Popular in multiple disciplines.
- Popular in industry.

# Interactive Data Language (IDL)

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

```
~  
pro helloworld.pro  
print, 'Hello World!'  
end  
~  
~  
~  
~  
~  
~
```



(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

The screenshot shows Microsoft SQL Server Management Studio with a query window open. The query is a SELECT statement with several columns and a FROM clause. The results window shows a table with 8 rows and 7 columns. The columns are labeled with '(No column name)' and 'petro u inclincorr'. The results are as follows:

(No column name)	(No column name)	(No column name)	(No column name)	(No column name)	(No column name)	petro u inclincorr	peti
1	17.74668	16.69974	16.35954	16.21339	16.35263	0.492397508119744	0.3
2	18.43394	17.29214	16.62328	16.28204	16.08086	1.52441919075319	0.9
3	19.42674	18.24301	17.70235	17.49629	17.15921	0.464193794342971	0.2
4	18.74544	17.94984	17.51807	17.26874	17.1523	0.500948764848324	0.3
5	18.949	17.80621	17.22119	16.86648	16.60248	0.68160486654221	0.4
6	19.22604	17.9269	17.52087	17.16631	17.18861	0.453664367877463	0.2
7	18.04691	16.87437	16.2979	16.01679	15.84145	0.837249857822637	0.5
8	18.11651	16.77158	16.13685	15.69099	15.65387	0.875719228372087	0.5

(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

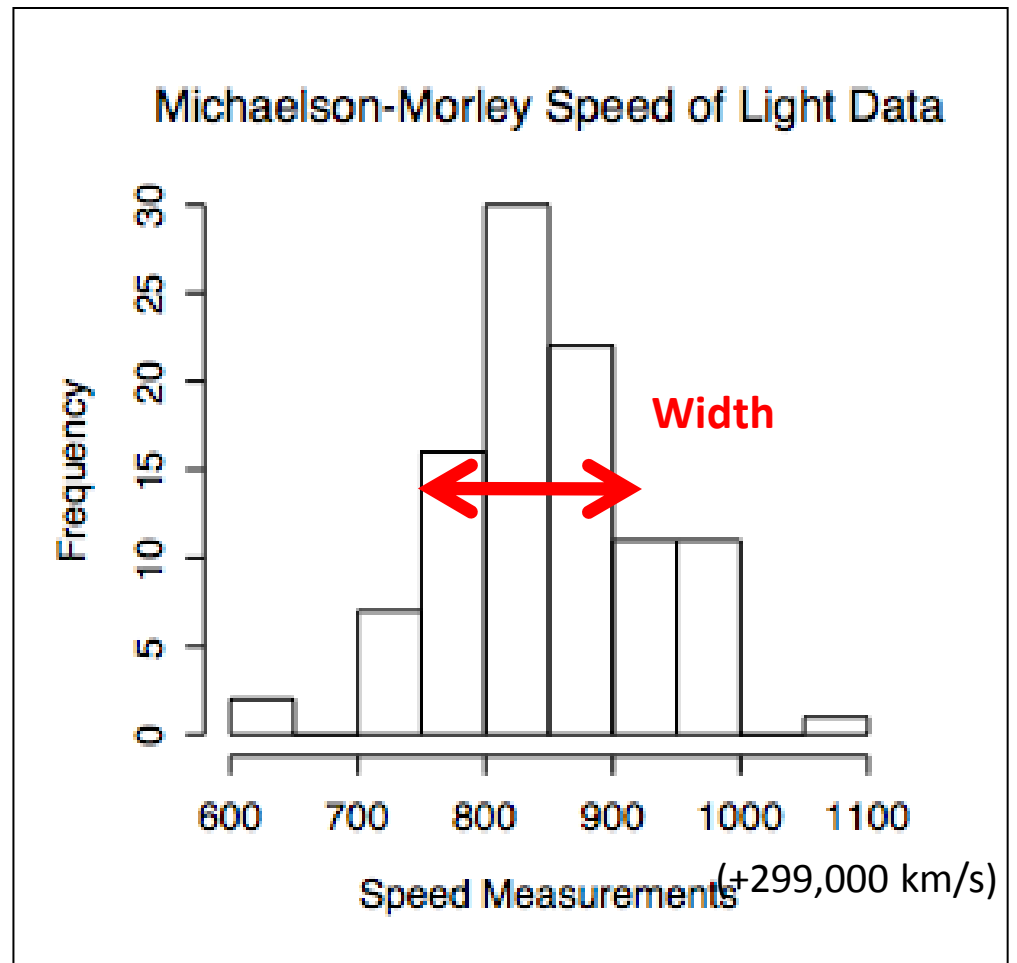
# Width of a Distribution: Standard Deviation (or “SD”, “ $\sigma$ ”)

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

$n$  = The number of data points

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

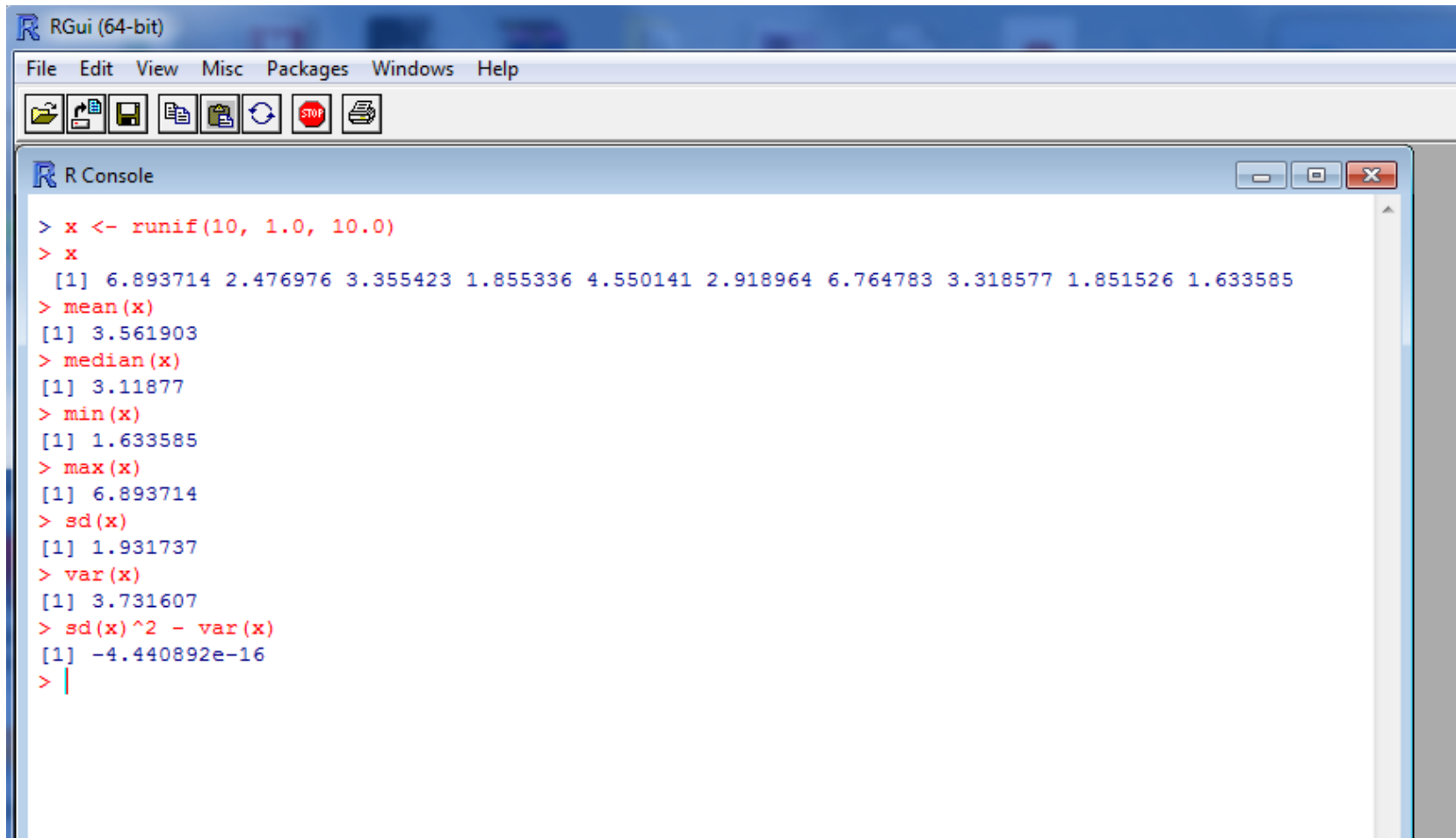
$x_i$  = Each of the values of the data



# 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



The screenshot shows the RGui (64-bit) window with a menu bar (File, Edit, View, Misc, Packages, Windows, Help) and a toolbar. The R Console window displays the following code and output:

```
> x <- runif(10, 1.0, 10.0)
> x
[1] 6.893714 2.476976 3.355423 1.855336 4.550141 2.918964 6.764783 3.318577 1.851526 1.633585
> mean(x)
[1] 3.561903
> median(x)
[1] 3.11877
> min(x)
[1] 1.633585
> max(x)
[1] 6.893714
> sd(x)
[1] 1.931737
> var(x)
[1] 3.731607
> sd(x)^2 - var(x)
[1] -4.440892e-16
> |
```

# 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*).

$X_1$        $X_2$        $X_3$        $X_4$      $\dots$        $X_{n-1}$      $X_n$

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*).

$X_1$        $X_2$        $X_3$        $X_4$      $\dots$        $X_{n-1}$      $X_n$

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, ...

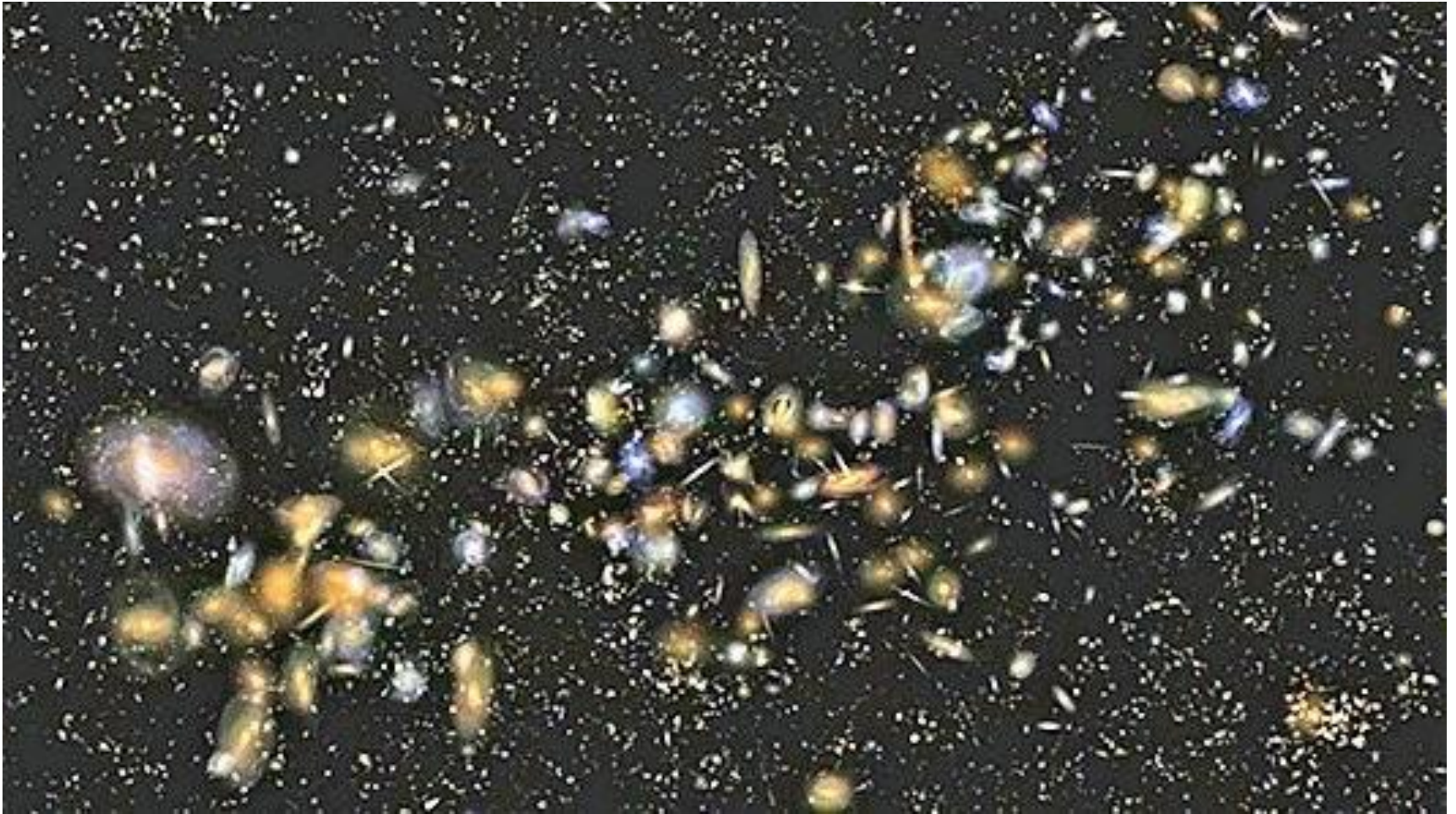
# Height of Trees in Forest



Sample  
Vs.  
Population



# Color of Galaxies in Universe

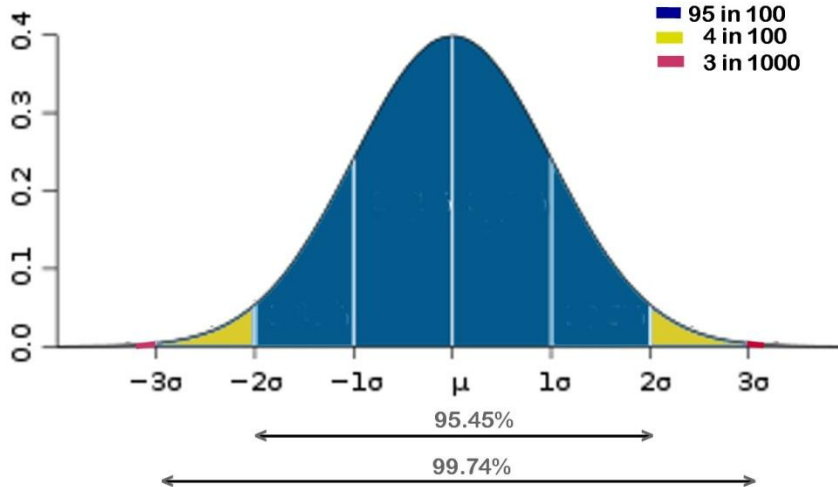


# Statistical Distributions

- 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)

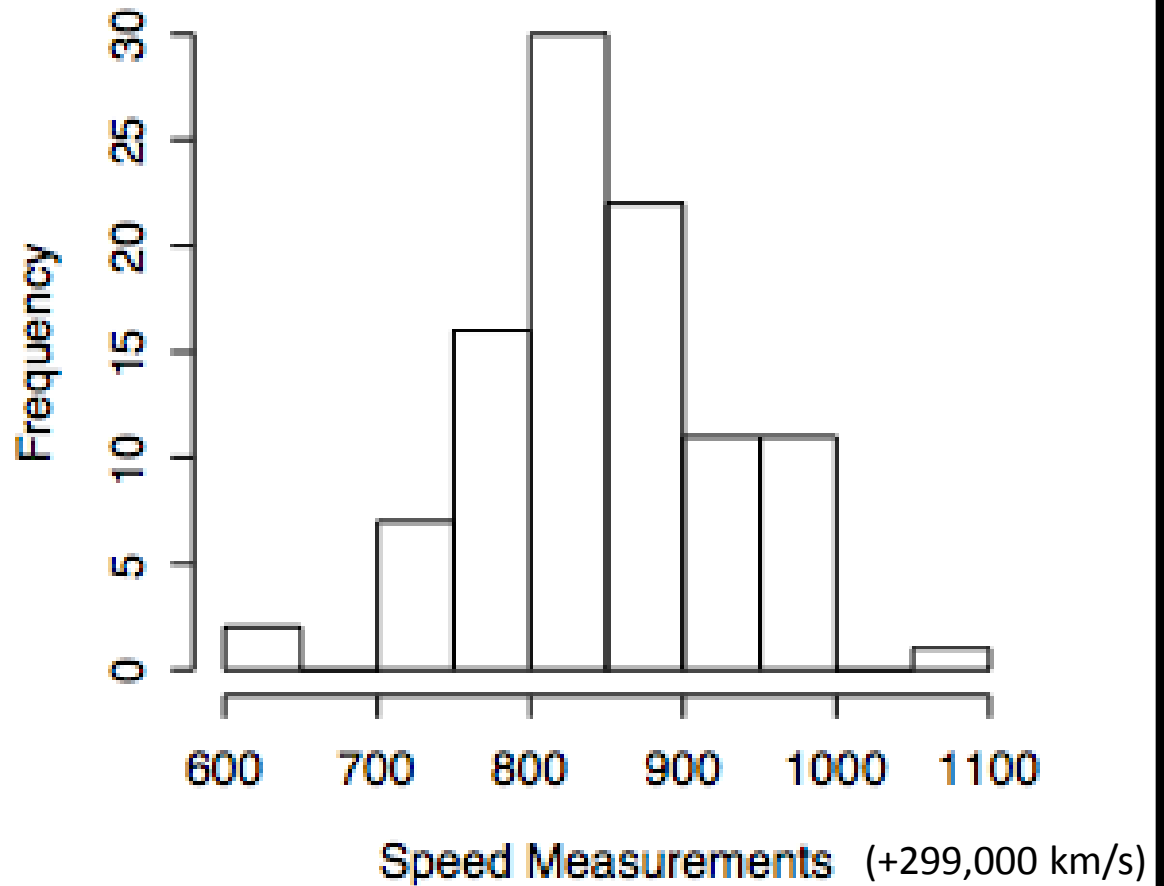
# Uncertainty in Physical Measurements

- 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

## Michaelson-Morley Speed of Light Data

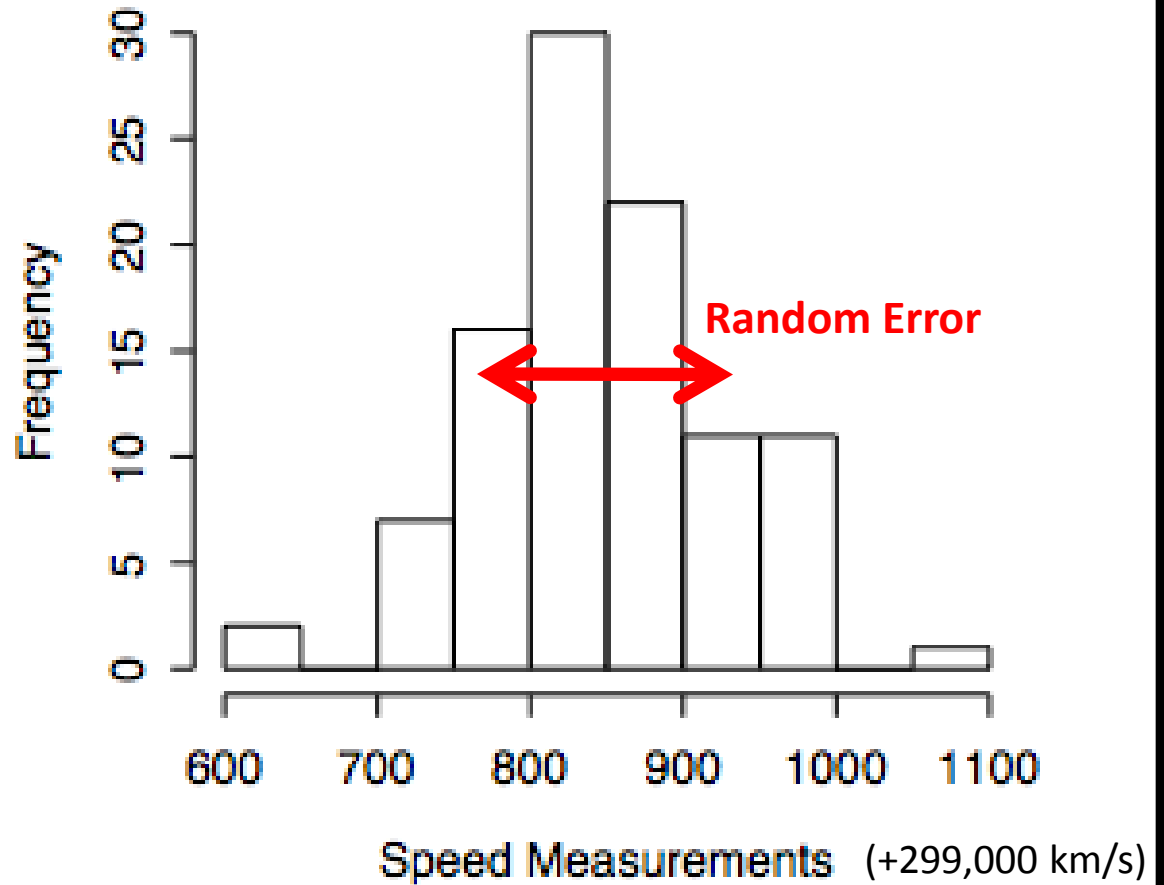






Albert Michelson at Chicago University  
Photo: University of Chicago

## Michaelson-Morley Speed of Light Data





# 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.*

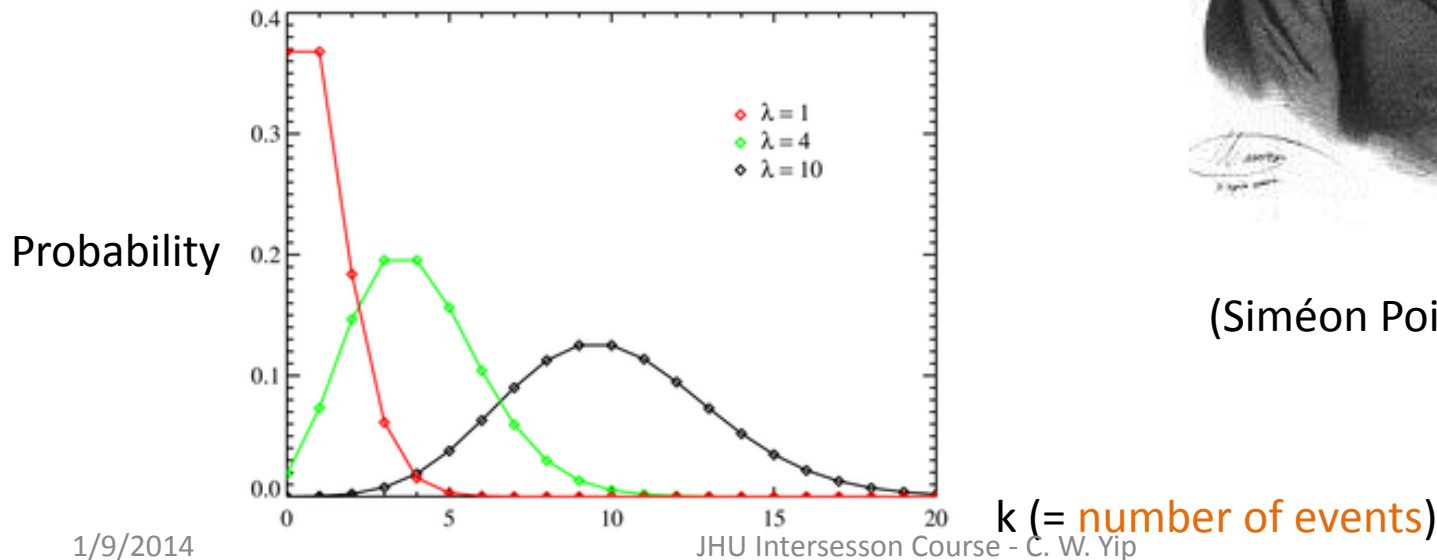
# Poisson Distribution

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

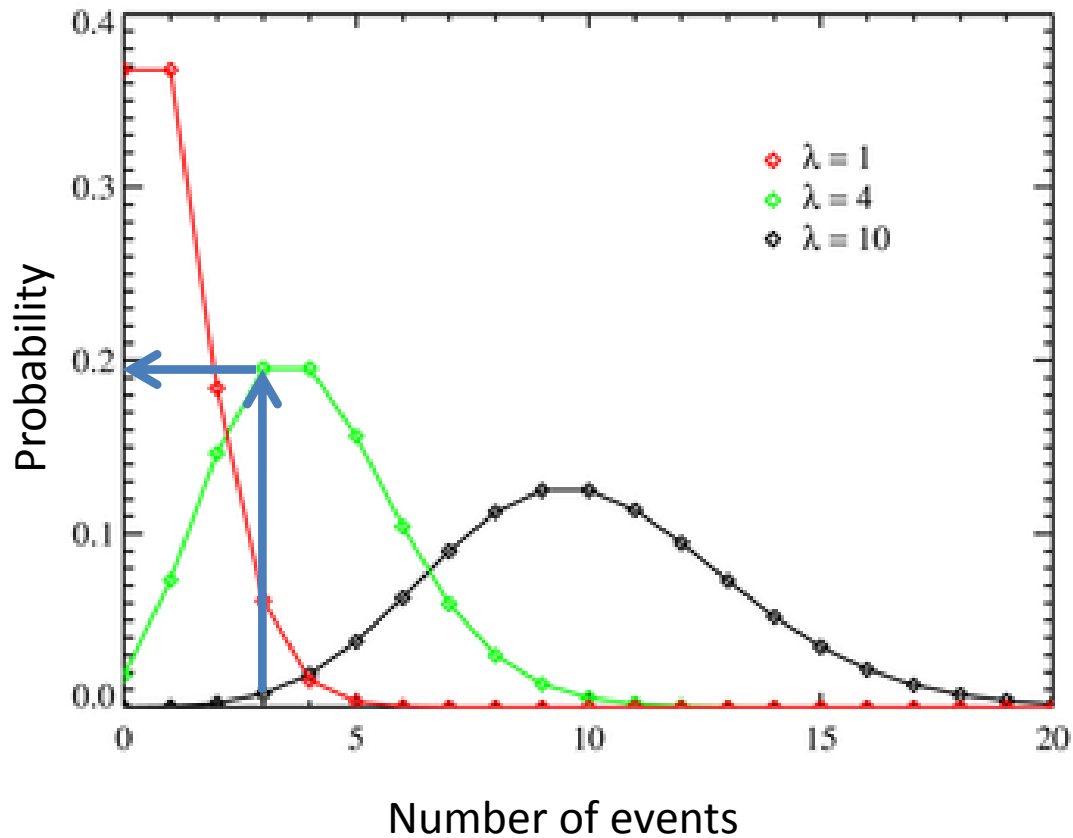
$$f(k; \lambda) = \frac{e^{-\lambda} \lambda^k}{k!},$$



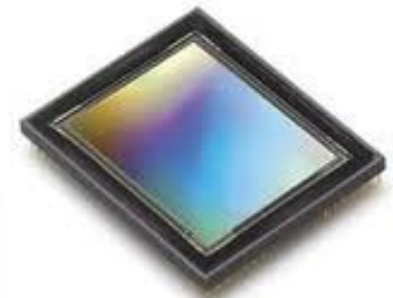
(Siméon Poisson, 1781-1840)



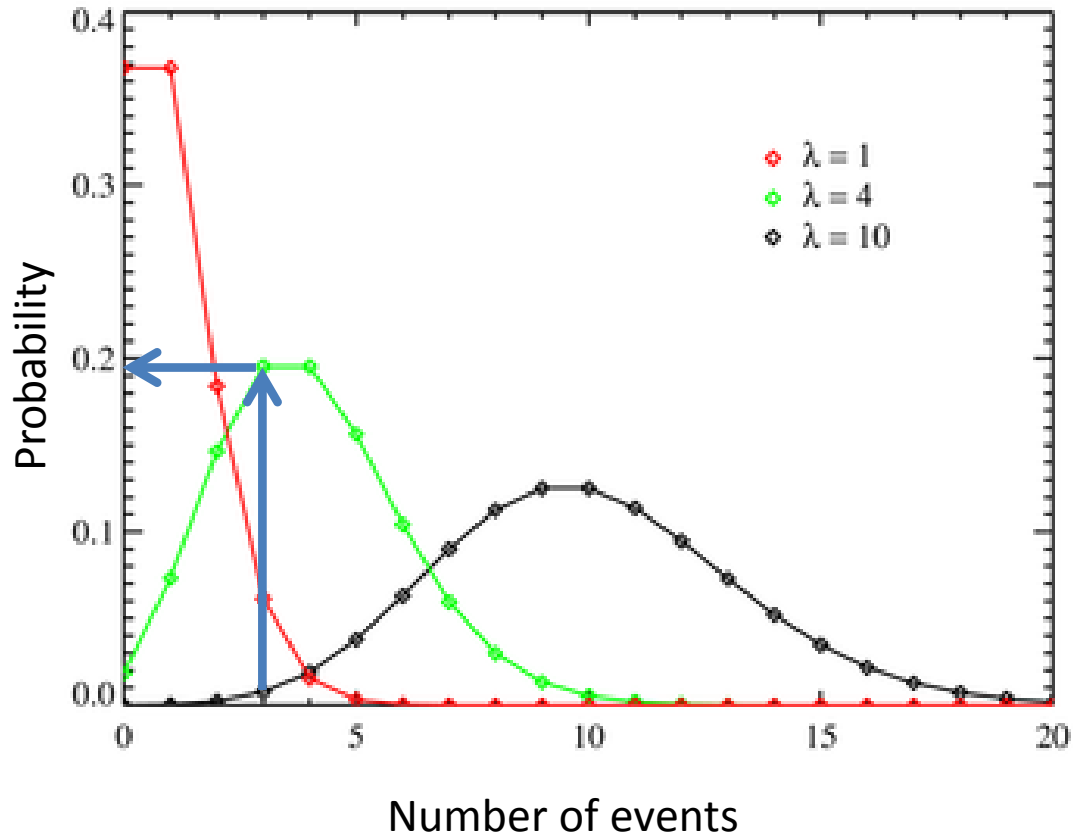
# Examples of Poisson Distributions



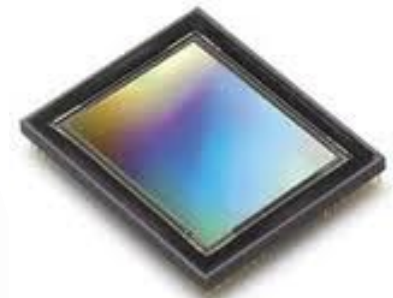
- I know from experience that I get 4 phone calls a day ( $\lambda = 4$ ).
- Then there is about 20% chance I will get 3 phone calls today.



# Examples of Poisson Distributions

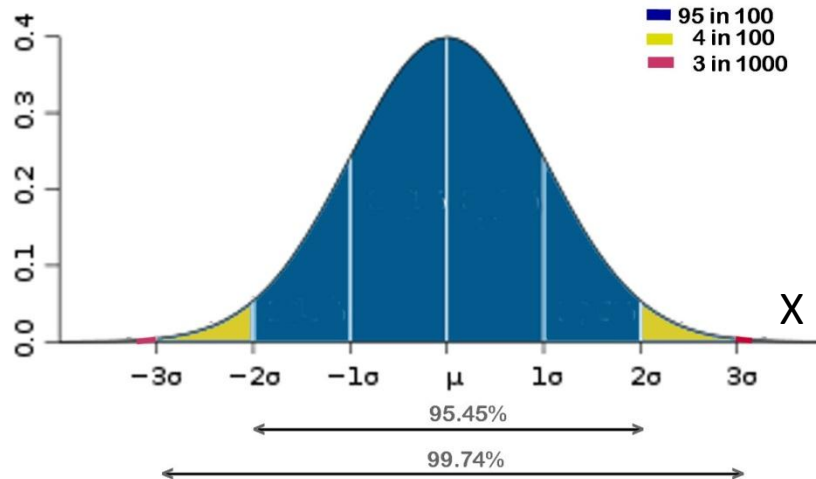


- I know from experience that I get 4 *photons* a day ( $\lambda = 4$ ).
- Then there is about 20% chance I will get 3 *photons* today.



# 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)

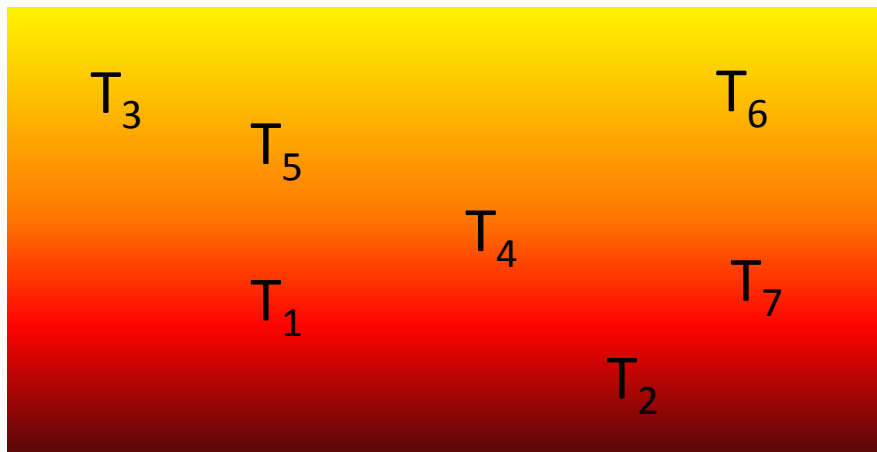


Probability of getting  $X$  between two values is the area under the Standard Normal Distribution between the two said values.

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

# Random Number: Gaussian Distribution as a Case Study

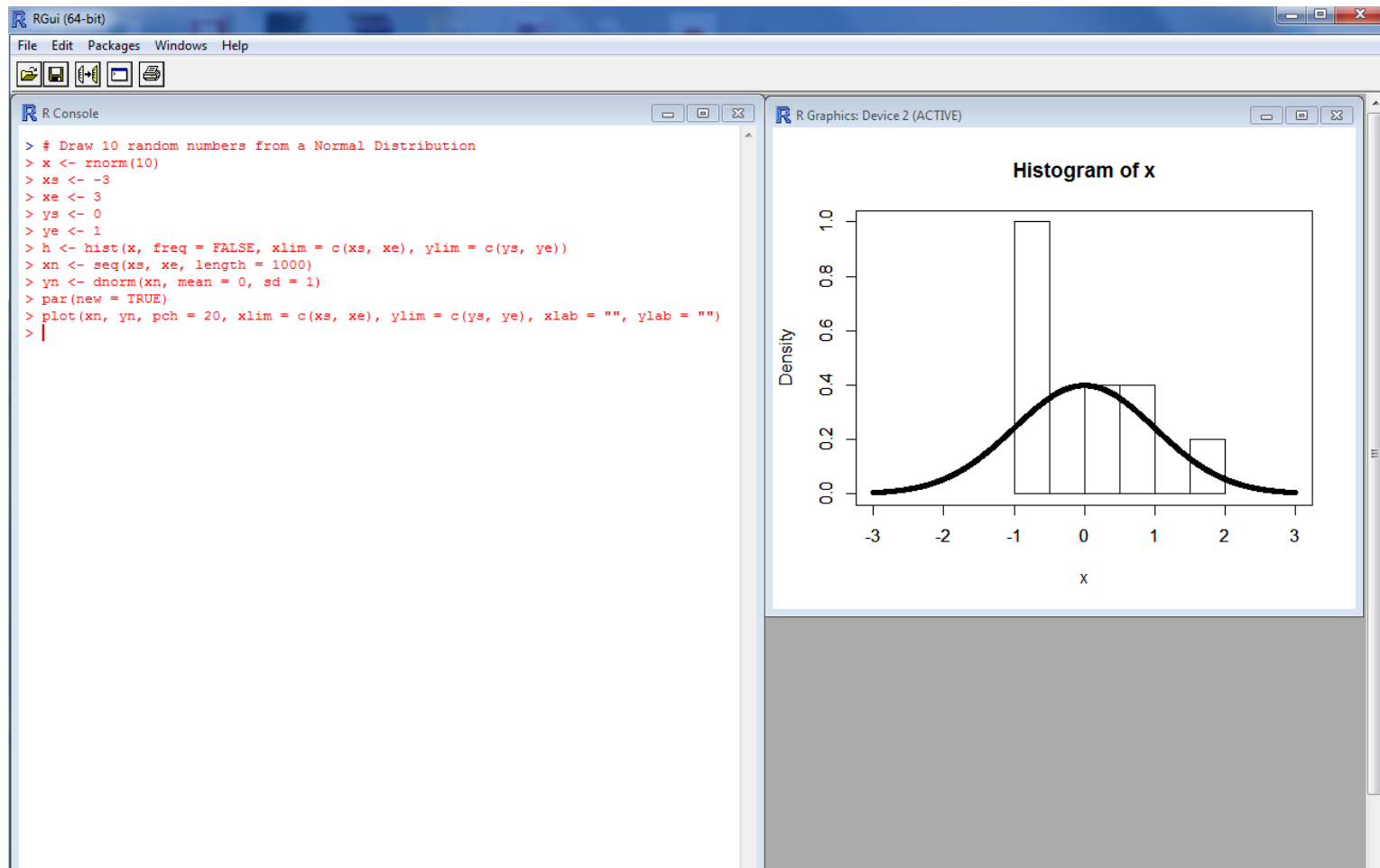
- When we analyze big datasets, *sampling methods* are commonly used.
- Random number plays an important role in sampling strategies.



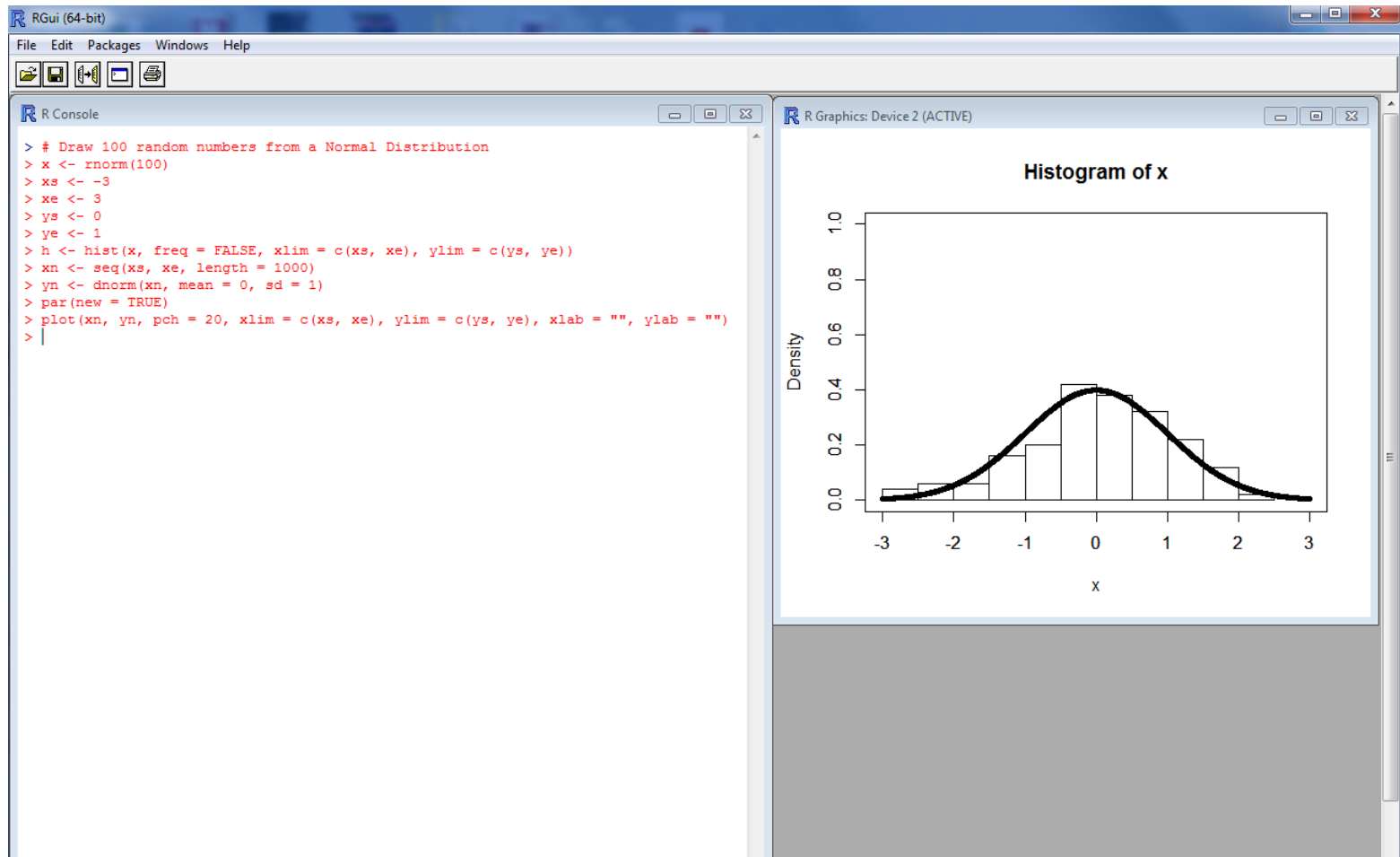
Suppose we have a piece of metal with temperature measurement  $T(x, y)$ .

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



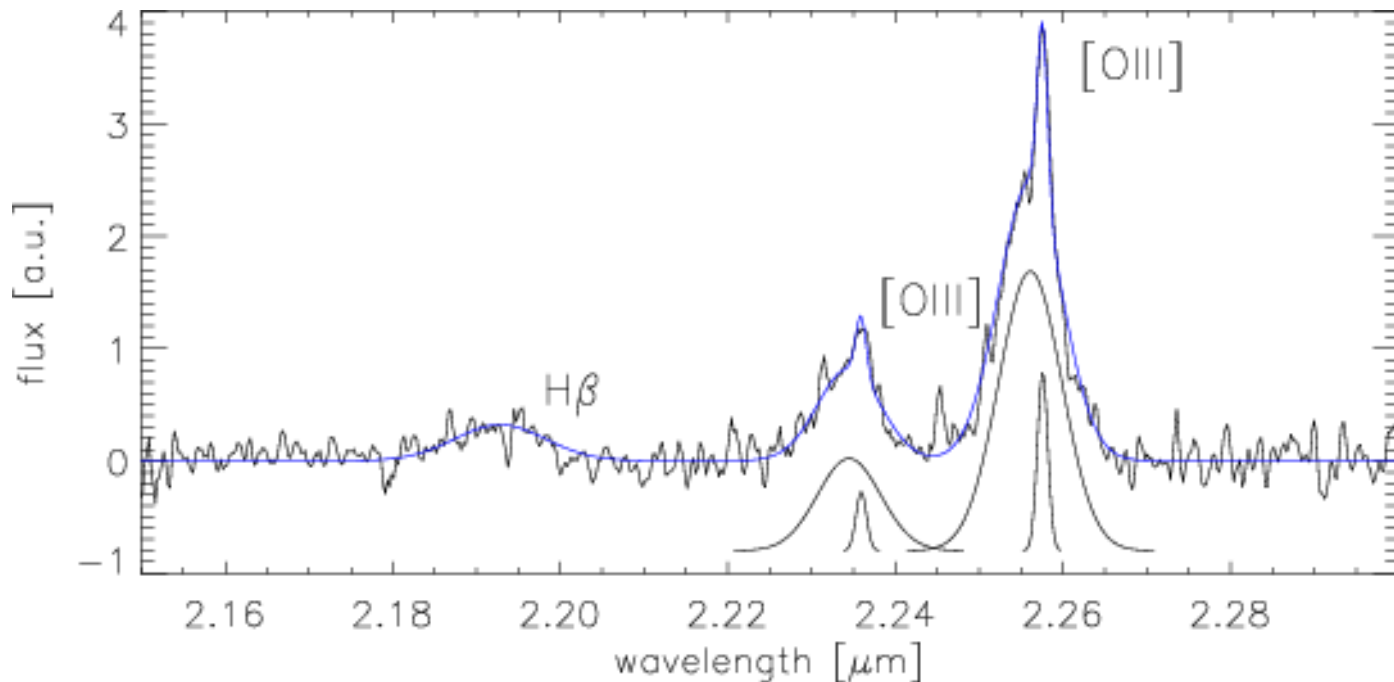


# Distributions in Astronomy

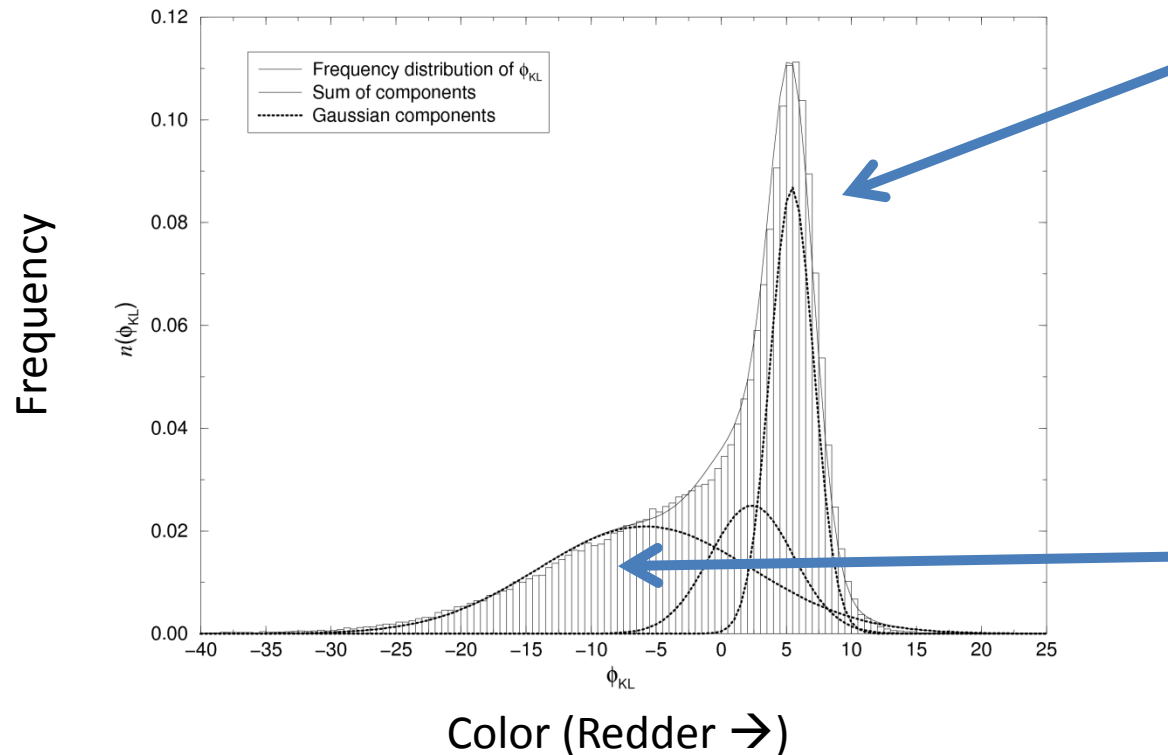
- Gaussian Distribution
- Poisson Distribution
- Planck Distribution
- Etc.

# Emission Lines in Galaxies: Gaussian

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

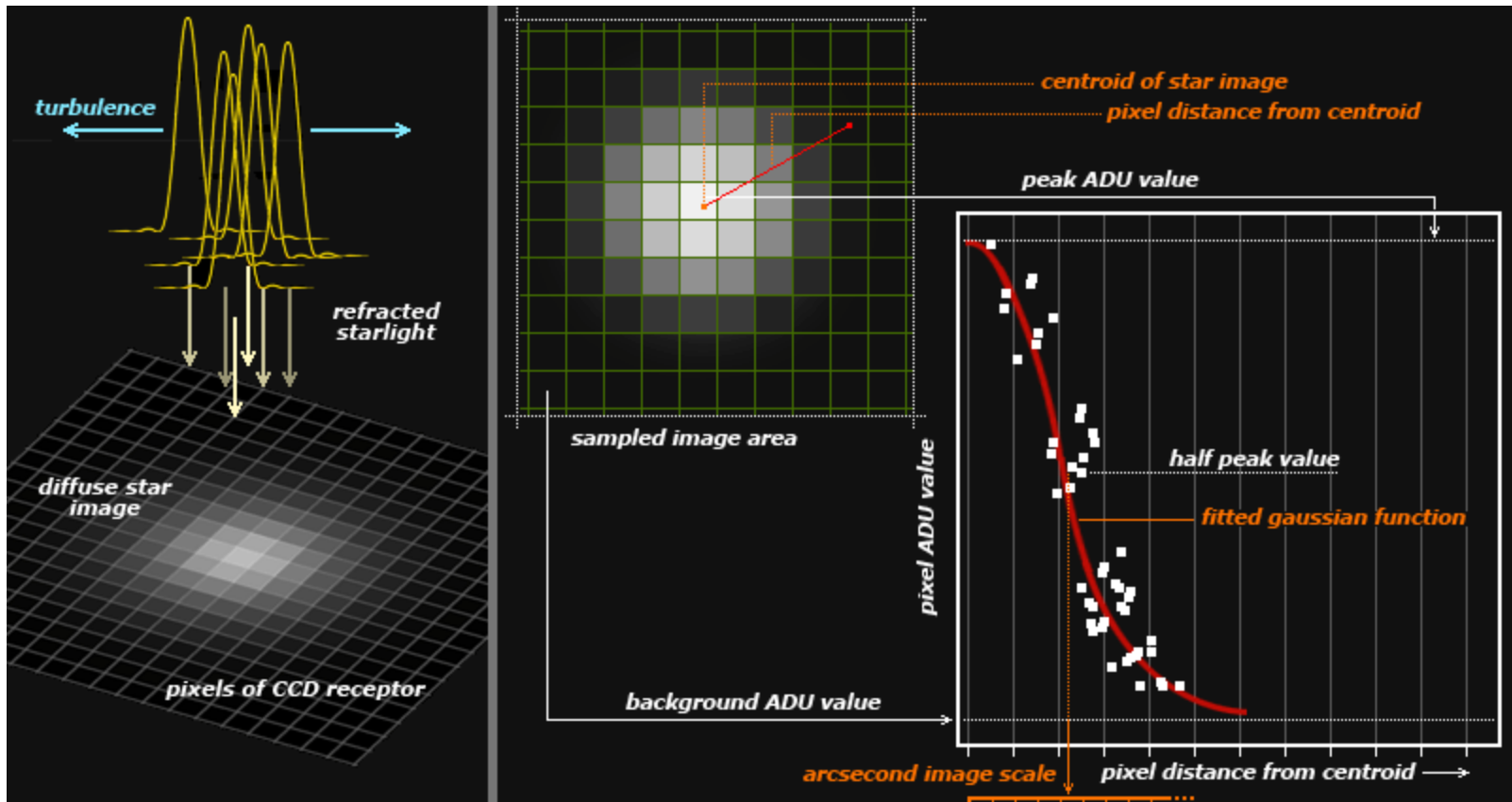


# Colors of Galaxies in Nearby Universe: Multiple Gaussians



(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

Chapter 2: Counting Photons

Figure 2.2 The greater the number of photons, the better the signal-to-noise ratio. In this case blurry, the photon count across the face of the galaxy image varies from a mean value of 0.1 photons to a mean value of 10,000 photons. The signal-to-noise ratio equals the square root of the photon count.

36 Handbook of Astronomical Image Processing

Section 2.3: Signals and Noise

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (\text{Eqn. 2.2})$$

Even after we have summed  $n$  samples, however, we still don't know the exact value of  $\bar{x}$  because the mean photon count remains uncertain by  $\sqrt{\bar{x}}$ .

Suppose that instead of examining the same pixel in a sequence of images, we were to examine one image in an area of sky that has no stars, so each you or pixels that are side-by-side is an independent sample of sky brightness, and it obeys the same rule that a series of samples at the same location does.

This has a profound impact on the collection of astronomical data. If you take two "identical" images of the same object and compare them, you will find they are not identical. The variation in the 100-photon signal is twice that of the 25-photon signal, does that mean the 100-photon signal is worse? In one sense it is—it has twice the standard deviation. However, the percentage variation in the 25-photon signal is  $5/25$ , or 20%, while in the 100-photon signal, the percentage variation is  $10/100$ , or 10%. Even though it has twice the standard deviation, the 100-photon signal has only half the percentage variation.

To quantify signal quality, engineers invented the signal-to-noise ratio (SNR):

$$\text{SNR} = \frac{\bar{x}}{\sqrt{\bar{x}}} = \sqrt{\bar{x}} \quad (\text{Eqn. 2.3})$$

The SNR of the 25-photon signal above is 5, and the SNR of the 100-photon signal is 10. The greater the signal-to-noise ratio, the better the image quality. Note, however, that within a given image, the signal is not the same at all places. The sky background will have a lower signal level than the bright center of a galaxy, so it is meaningless to assign an SNR to an entire image because signal-to-noise ratio is meaningful for only one signal level. Nevertheless, astronomers sometimes quote an SNR for an image, and when they do, it refers to the SNR at the signal level of the sky background.

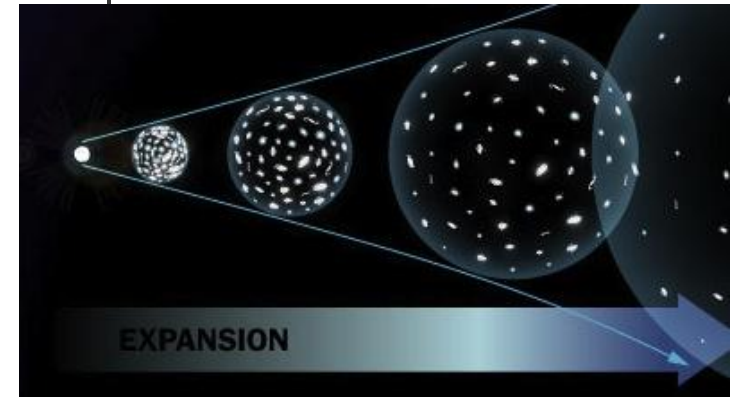
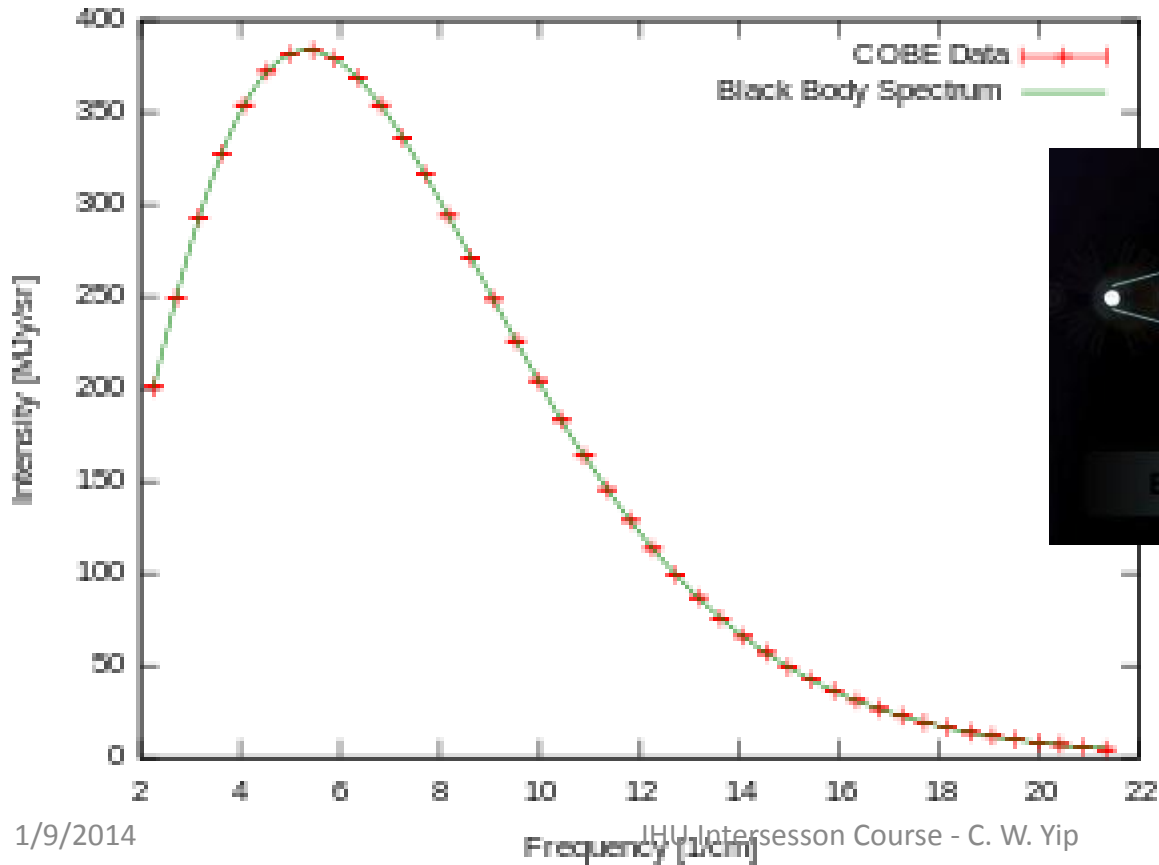
Richard Berry and James Burnell 37

**Signal-to-Noise Ratio (SNR)**  
 = Count/Random Error of Count  
 =  $\sqrt{\text{Count}}$  (for Poisson)  
 That is, when Count increases, SNR increases.

# Cosmic Microwave Background: Planck Distribution

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

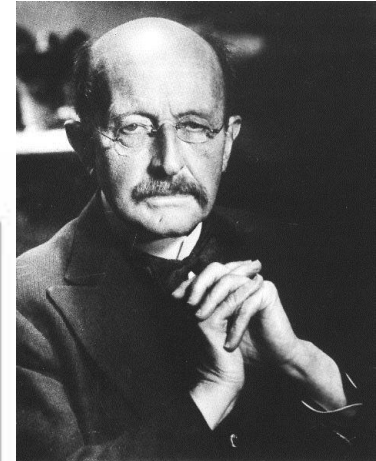
Cosmic Microwave Background Spectrum from COBE



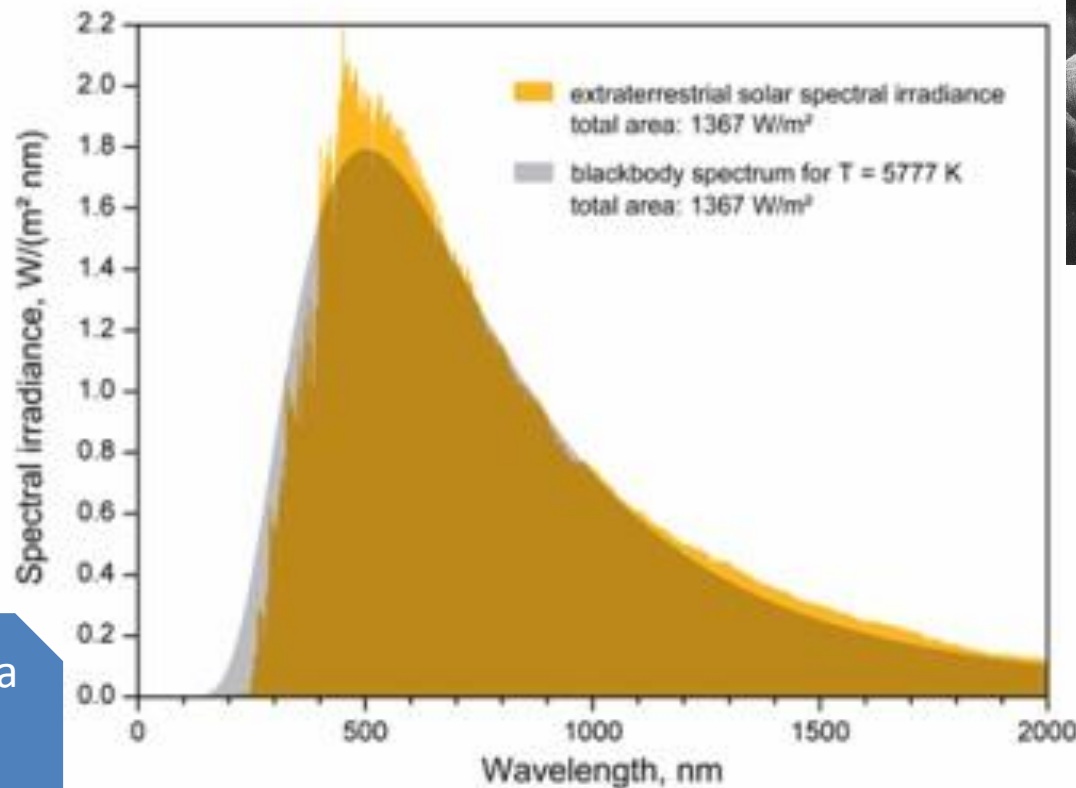
- Universe expands and cools down.
- The photons were hot, but cold now (3 Kelvin).



# Black Body Radiation from the Sun: Planck Distribution



(Max Planck,  
1858 – 1947)



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