

Data Mining In Modern Astronomy Sky Surveys:

*Hypothesis Testing,
Bayes' Theorem,
& Parameter Estimation*

Ching-Wa Yip

cwyip@pha.jhu.edu; **Bloomberg 518**

Erratum of Last Lecture

- The Central Limit Theorem was proved by Bernoulli back in 17th century. The Michelson-Morley speed-of-light experiment was carried out in 18th century.
- Michelson & Morley could have accessed to the Central Limit Theorem and decided to carry out many, repeated measurements. (Need more research)

From Data to Information

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

Information



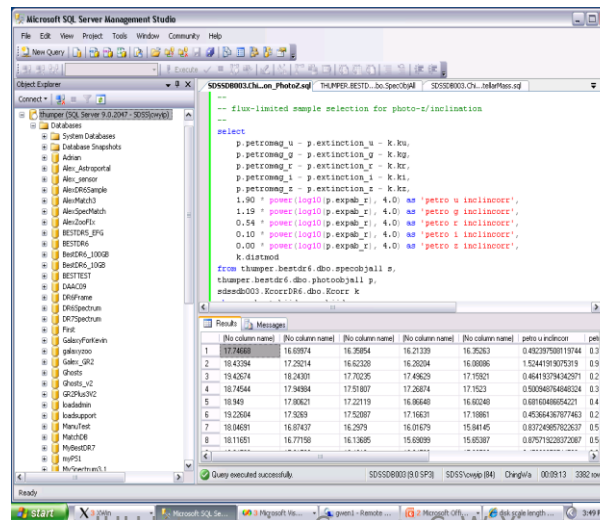
Database



Sensors



Data Analysis
or
Data Mining



From Data to Information

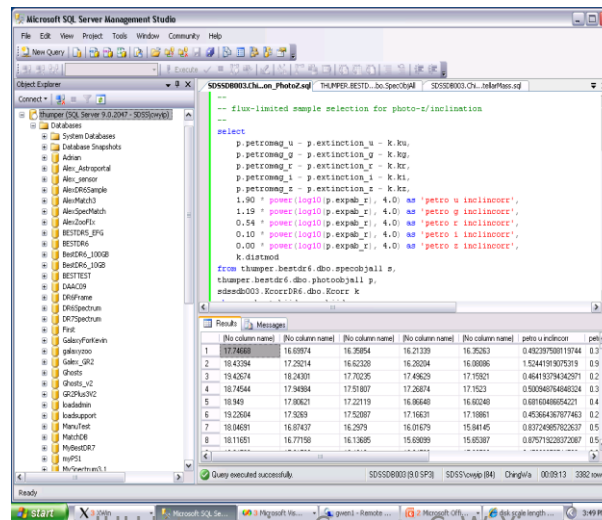
- We don't just want data.
- We want information from the data.

Information

Database

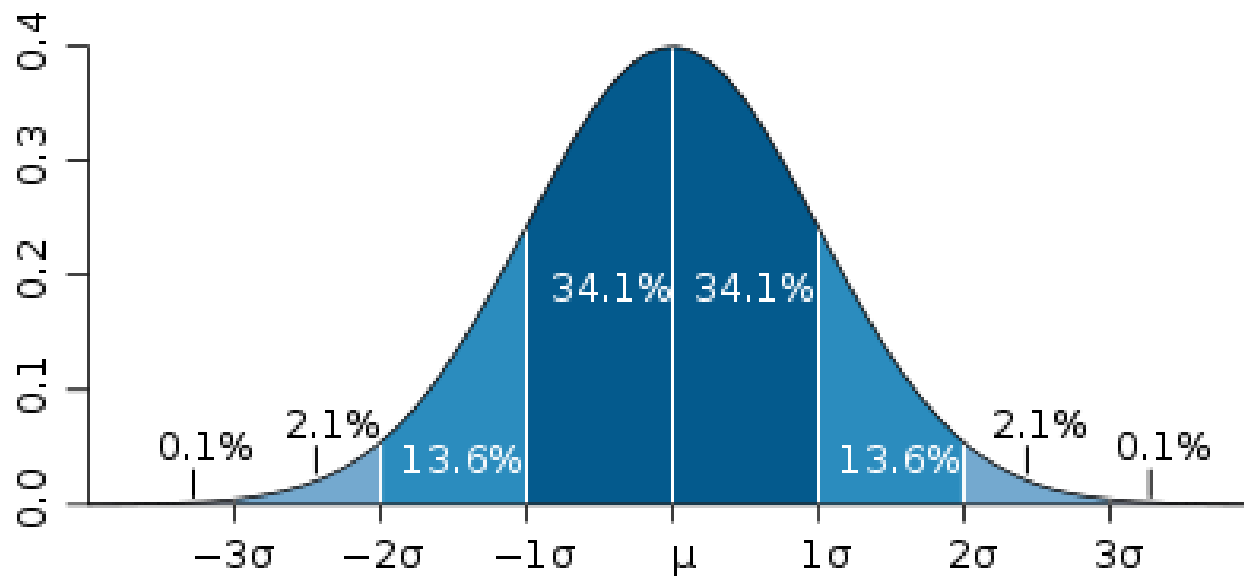
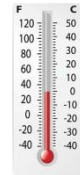
Sensors

Data Analysis
or
Data Mining



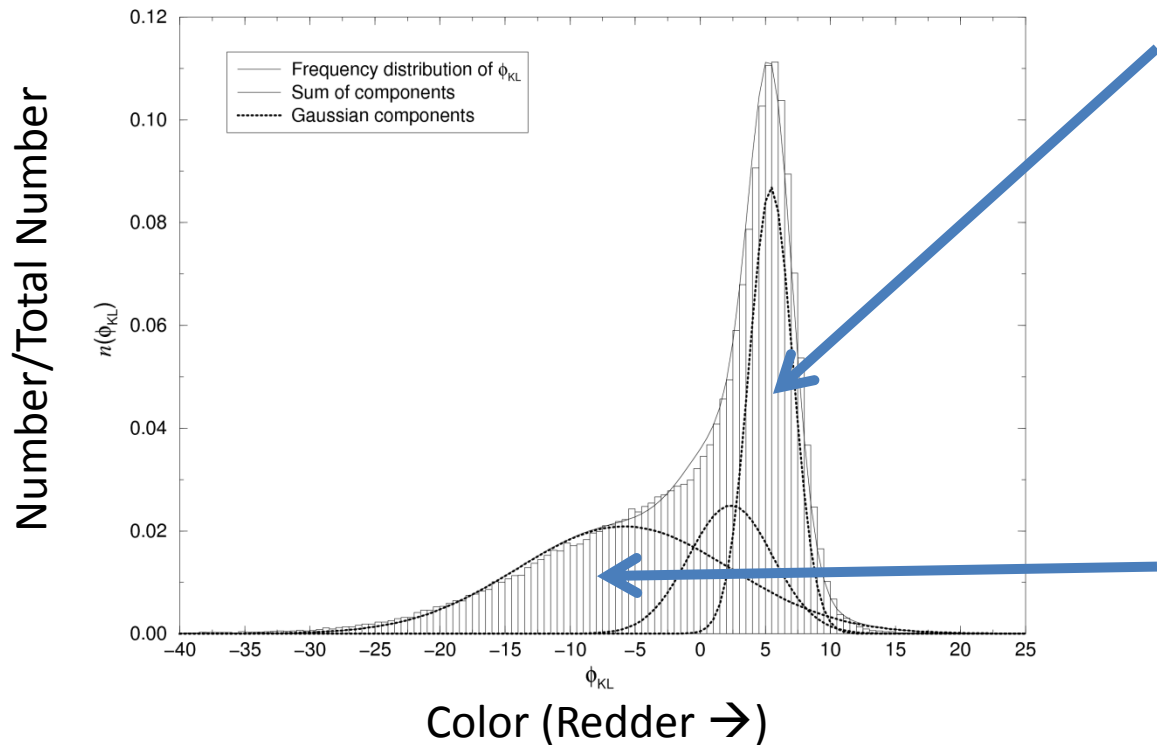
Probability Density Function (PDF)

- A function which tells the probability of an event, e.g.:
 - Temperature T lies between 34F and 50F.
 - Variable X lies between X_1 and X_2 .
- A well-known PDF is the Standard Normal Distribution:



PDFs Come in Many Shapes

- E.g., Color of galaxies



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

Properties of PDFs

- The total area under the function is 1 (= 100% probability):

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

- The probability of the variable x lying between a and b is:

$$P(x \text{ between } a \text{ and } b) = \int_a^b p(x)dx$$

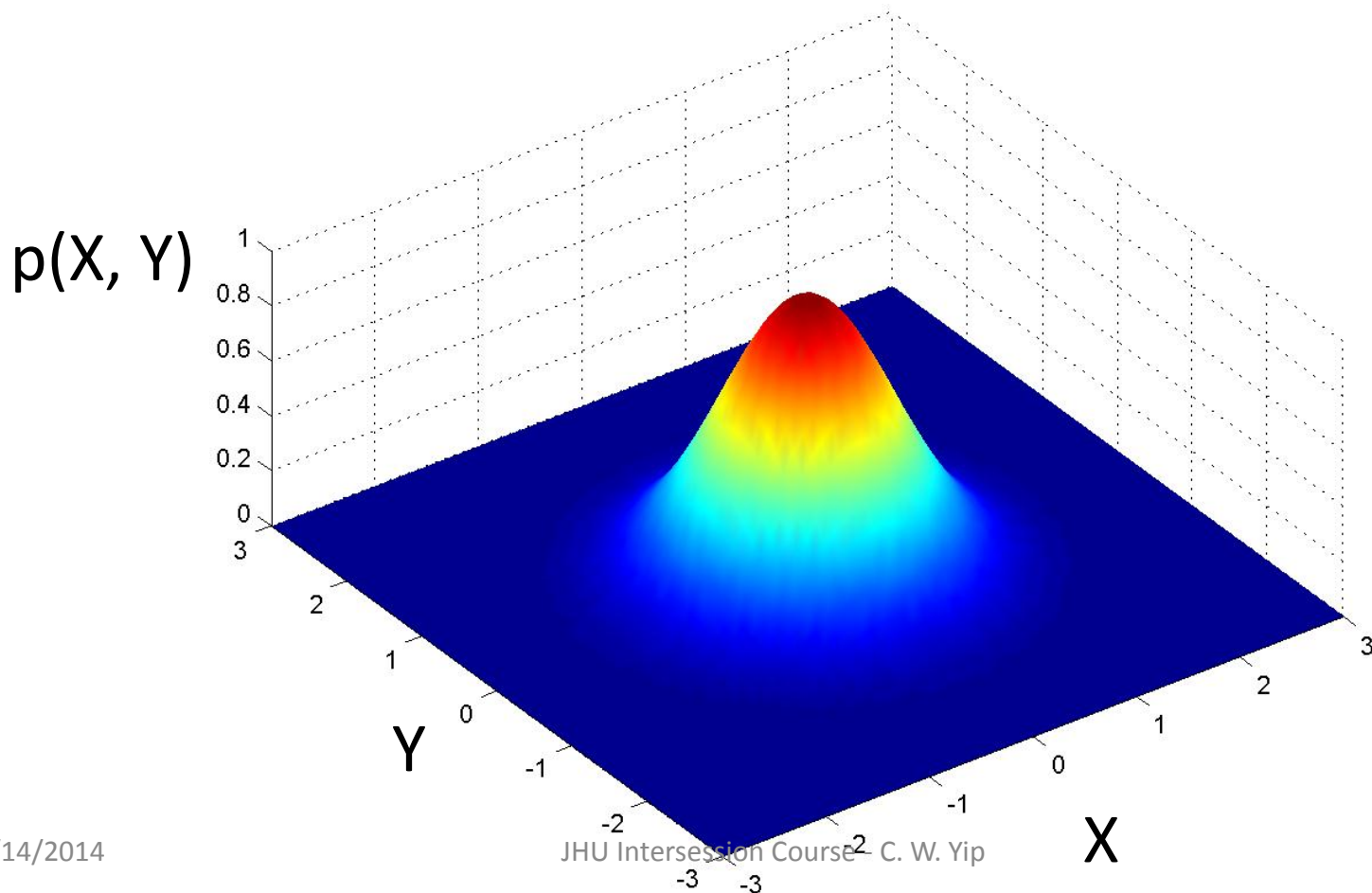
2D Probability Density Function (PDF)

- A function which tells the probability of an event, just like the 1D PDF but with 2 variables:
 - the variables X lies between X_1 and X_2 **AND Y** **between Y_1 and Y_2**
- The total area under the curve is still 1.

$$\iint_{-\infty}^{\infty} p(x, y) dx dy = 1$$

Example 2D PDF

- Seeing disk of stellar image.



Standard Normal Distribution Revisit: Some Terminologies

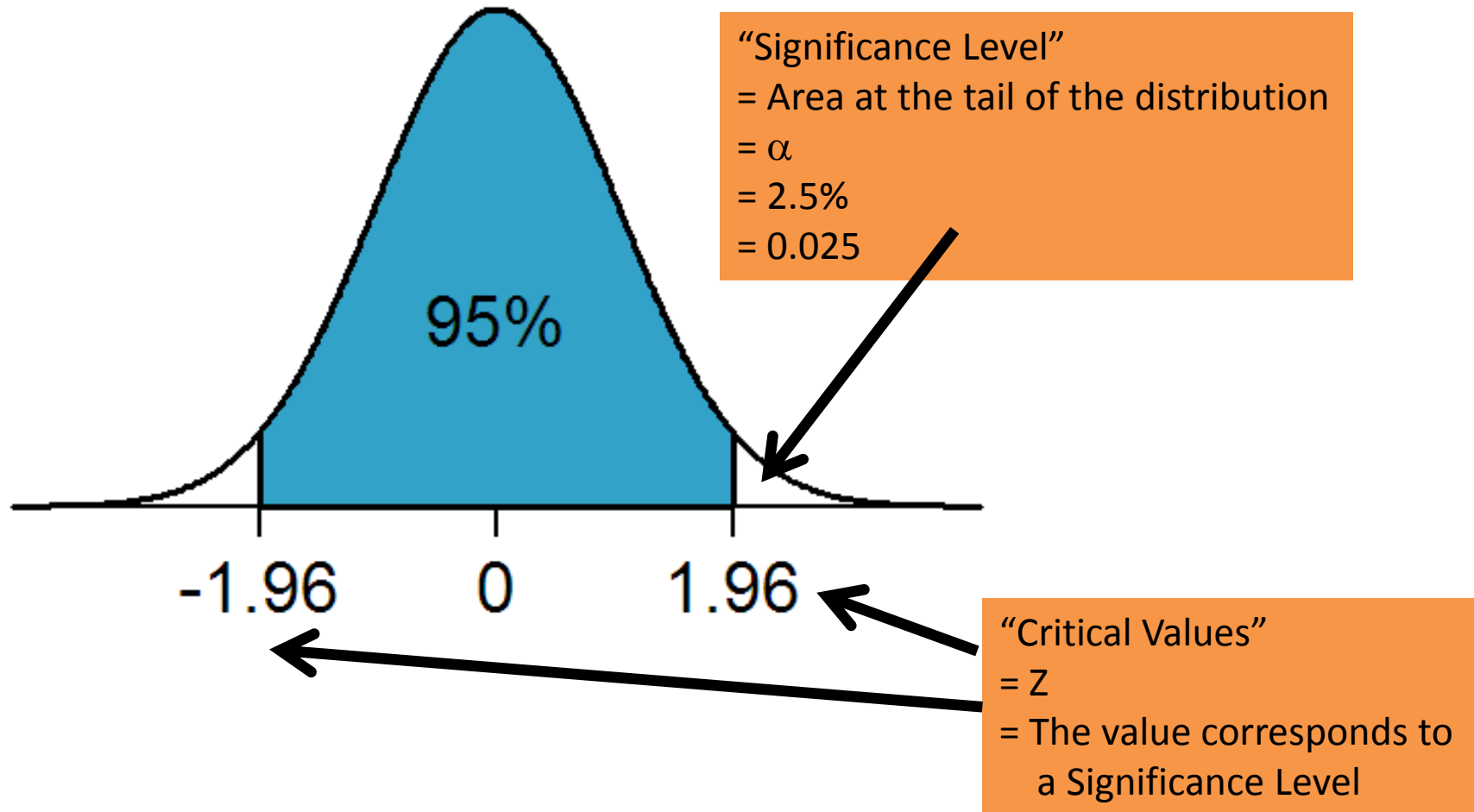


Table B.1. Right tail probabilities $1 - \Phi(a) = P(Z \geq a)$ for an $N(0, 1)$ distributed random variable Z .

a	0	1	2	3	4	5	6	7	8	9
0.0	5000	4960	4920	4880	4840	4801	4761	4721	4681	4641
0.1	4602	4562	4522	4483	4443	4404	4364	4325	4286	4247
0.2	4207	4168	4129	4090	4052	4013	3974	3936	3897	3859
0.3	3821	3783	3745	3707	3669	3632	3594	3557	3520	3483
0.4	3446	3409	3372	3336	3300	3264	3228	3192	3156	3121
0.5	3085	3050	3015	2981	2946	2912	2877	2843	2810	2776
0.6	2743	2709	2676	2643	2611	2578	2546	2514	2483	2451
0.7	2420	2389	2358	2327	2296	2266	2236	2206	2177	2148
0.8	2119	2090	2061	2033	2005	1977	1949	1922	1894	1867
0.9	1841	1814	1788	1762	1736	1711	1685	1660	1635	1611
1.0	1587	1562	1539	1515	1492	1469	1446	1423	1401	1379
1.1	1357	1335	1314	1292	1271	1251	1230	1210	1190	1170
1.2	1151	1131	1112	1093	1075	1056	1038	1020	1003	0985
1.3	0968	0951	0934	0918	0901	0885	0869	0853	0838	0823
1.4	0808	0793	0778	0764	0749	0735	0721	0708	0694	0681
1.5	0668	0655	0643	0630	0618	0606	0594	0582	0571	0559
1.6	0548	0537	0526	0516	0505	0495	0485	0475	0465	0455
1.7	0446	0436	0427	0418	0409	0401	0392	0384	0375	0367
1.8	0359	0351	0344	0336	0329	0322	0314	0307	0301	0294
1.9	0287	0281	0274	0268	0262	0256	0250	0244	0239	0233
2.0	0228	0222	0217	0212	0207	0202	0197	0192	0188	0183
2.1	0179	0174	0170	0166	0162	0158	0154	0150	0146	0143
2.2	0139	0136	0132	0129	0125	0122	0119	0116	0113	0110
2.3	0107	0104	0102	0099	0096	0094	0091	0089	0087	0084
2.4	0082	0080	0078	0075	0073	0071	0069	0068	0066	0064

(Abridged,
Taken from
Dekking et al.)

Table B.1. Right tail probabilities $1 - \Phi(a) = P(Z \geq a)$ for an $N(0, 1)$ distributed random variable Z .

a	0	1	2	3	4	5	6	7	8	9
0.0	5000	4960	4920	4880	4840	4801	4761	4721	4681	4641
0.1	4602	4562	4522	4483	4443	4404	4364	4325	4286	4247
0.2	4207	4168	4129	4090	4052	4013	3974	3936	3897	3859
0.3	3821	3782	3745	3707	3669	3632	3594	3557	3520	3483
0.4	3445	3408	3371	3334	3297	3261	3228	3192	3156	3121
0.5	3078	3042	3006	2971	2936	2901	2877	2843	2810	2776
0.6	2719	2684	2649	2615	2581	2546	2514	2483	2451	2421
0.7	2566	2532	2498	2465	2432	2399	2366	2334	2302	2271
0.8	2311	2279	2247	2215	2183	2151	2120	2089	2058	2028
0.9	1841	1814	1788	1762	1736	1711	1685	1660	1635	1611
1.0	1587	1562	1539	1515	1492	1469	1446	1423	1401	1379
1.1	1357	1335	1314	1292	1271	1251	1230	1210	1190	1170
1.2	1151	1131	1112	1093	1075	1056	1038	1020	1003	0985
1.3	0968	0951	0934	0918	0901	0885	0869	0853	0838	0823
1.4	0808	0793	0778	0764	0749	0735	0721	0708	0694	0681
1.5	0668	0655	0643	0630	0618	0606	0594	0582	0571	0559
1.6	0548	0537	0526	0516	0505	0495	0485	0475	0465	0455
1.7	0446	0436	0427	0418	0409	0401	0392	0384	0375	0367
1.8	0359	0351	0344	0336	0329	0322	0314	0307	0301	0294
1.9	0287	0281	0274	0268	0262	0256	0250	0244	0239	0233
2.0	0228	0222	0217	0212	0207	0202	0197	0192	0188	0183
2.1	0179	0174	0170	0166	0162	0158	0154	0150	0146	0143
2.2	0139	0136	0132	0129	0125	0122	0119	0116	0113	0110
2.3	0107	0104	0102	0099	0096	0094	0091	0089	0087	0084
2.4	0082	0080	0078	0075	0073	0071	0069	0068	0066	0064

If:

$\alpha = 0.025$ (tabulated as 0250)

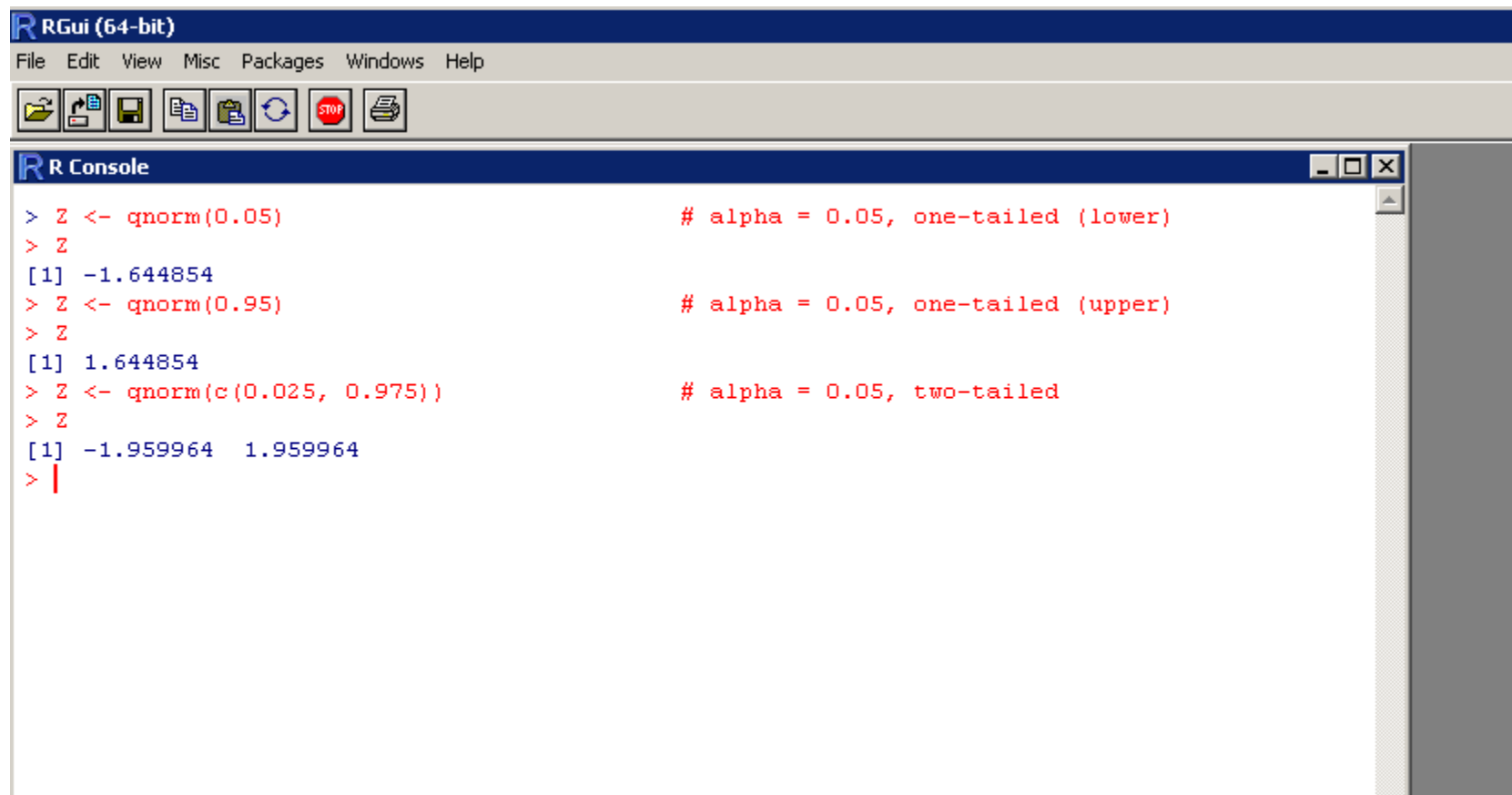
We get:

$Z = 1.96$

(Abridged,
Taken from
Dekking et al.)

Significance Level (α) and its Critical Values (Z)

- No more tables!



```
RGui (64-bit)
File Edit View Misc Packages Windows Help
[Icons: Home, Open, Save, Print, Refresh, Stop, Print]

R Console
> Z <- qnorm(0.05) # alpha = 0.05, one-tailed (lower)
> Z
[1] -1.644854
> Z <- qnorm(0.95) # alpha = 0.05, one-tailed (upper)
> Z
[1] 1.644854
> Z <- qnorm(c(0.025, 0.975)) # alpha = 0.05, two-tailed
> Z
[1] -1.959964 1.959964
> |
```

Hypothesis Testing

- Goal: To test whether a hypothesis is true or false.
- The beginning hypothesis is our best knowledge for the problem (also called the Null Hypothesis).
- If **Null Hypothesis** is FALSE, **Alternative Hypothesis** is TRUE.

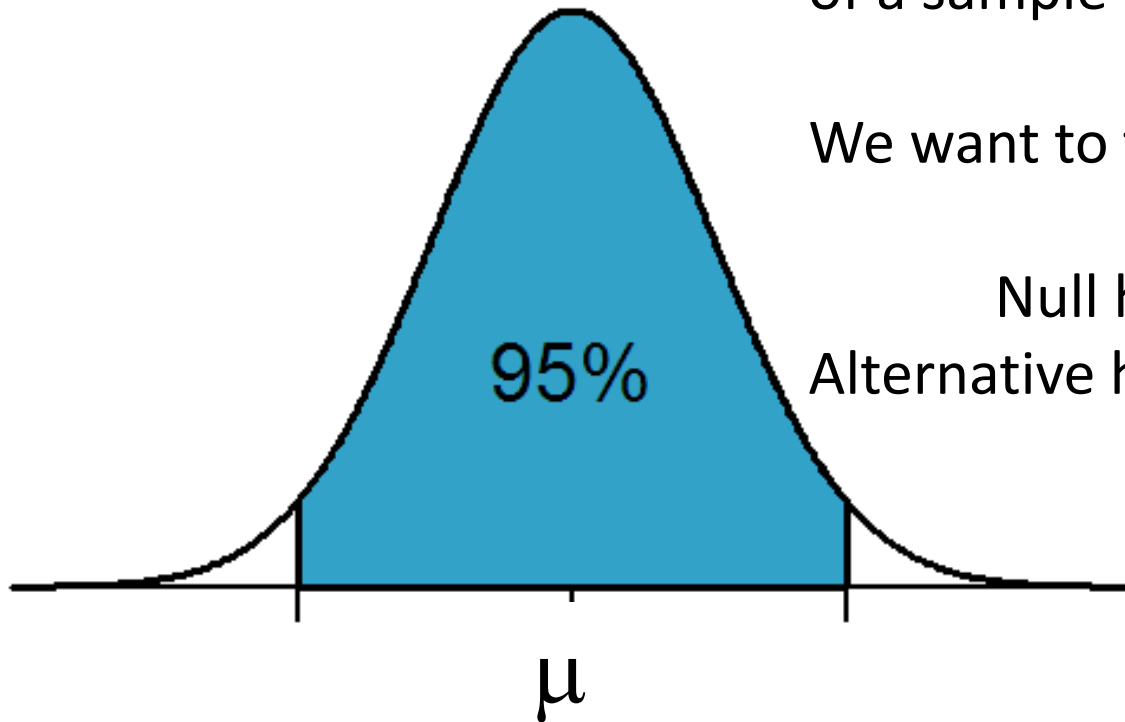
Hypothesis Testing

Suppose we think the average of a sample is some number.

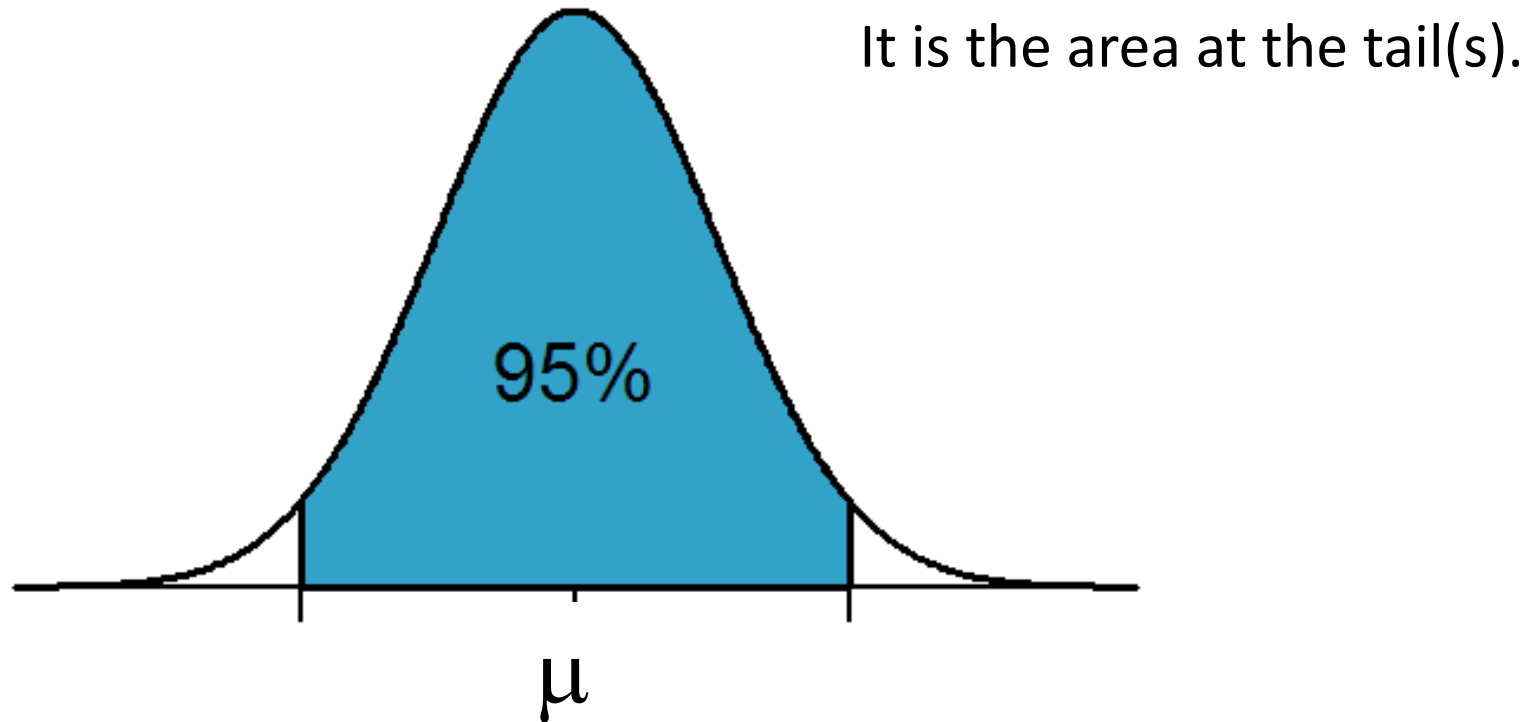
We want to test this hypothesis.

Null hypothesis: $\mu = \text{some \#}$

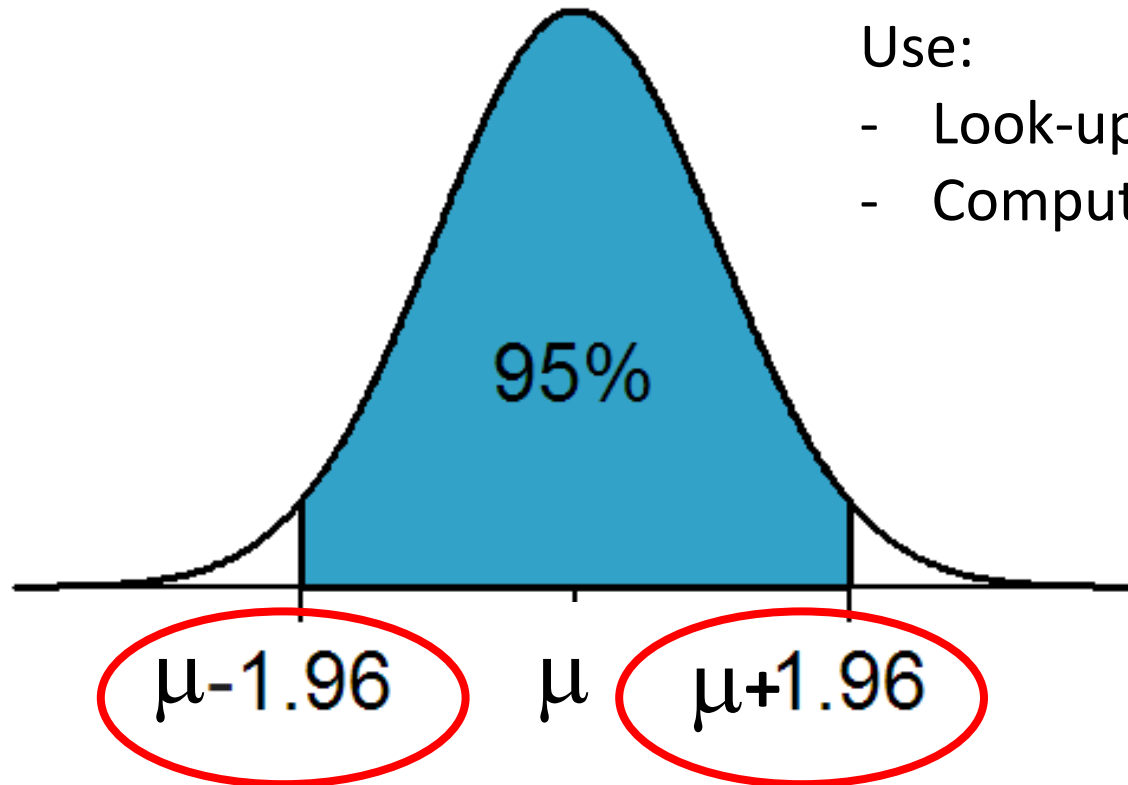
Alternative hypothesis: $\mu \neq \text{some \#}$



Hypothesis Testing: (Step 1) Set Significance Levels



Hypothesis Testing: (Step 2) Find Critical Values

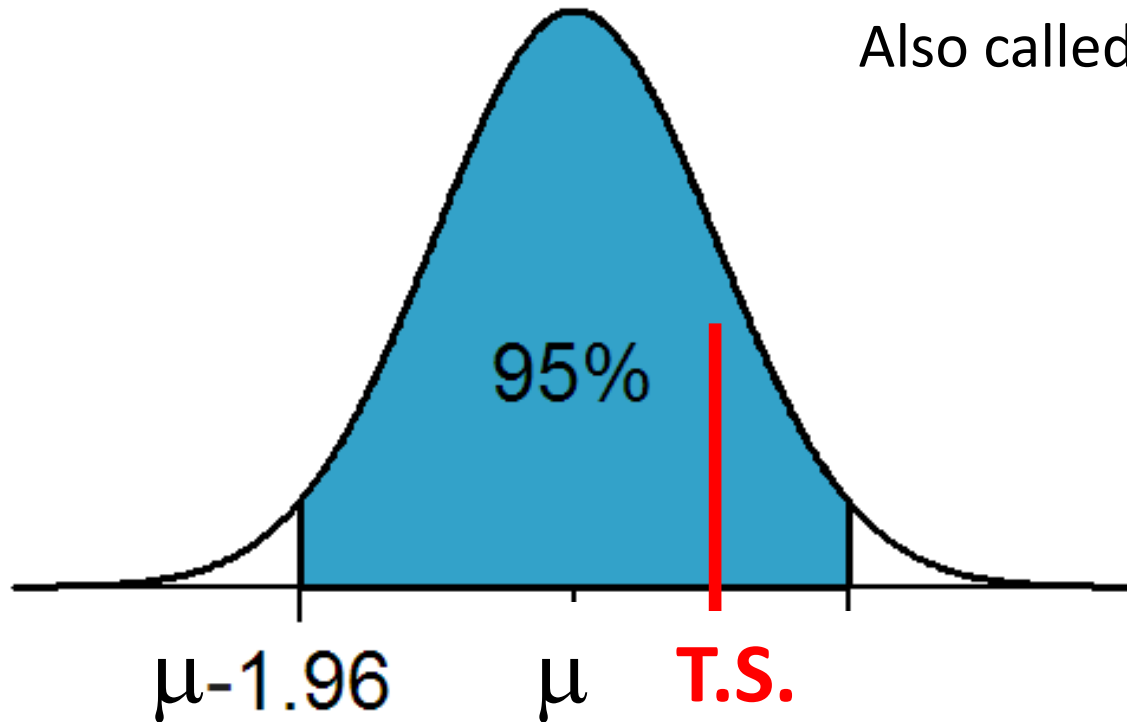


Use:

- Look-up tables
- Computational software

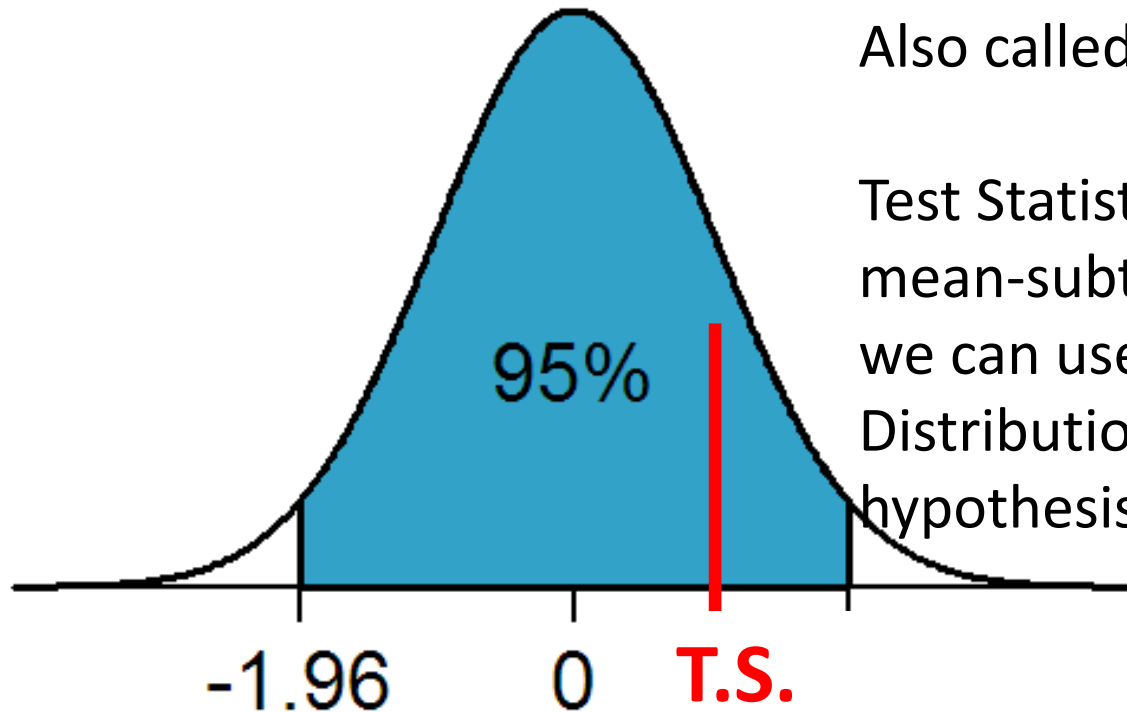
Hypothesis Testing: (Step 3) Calculate Statistics From Data

Also called **Test Statistics**.



Hypothesis Testing:

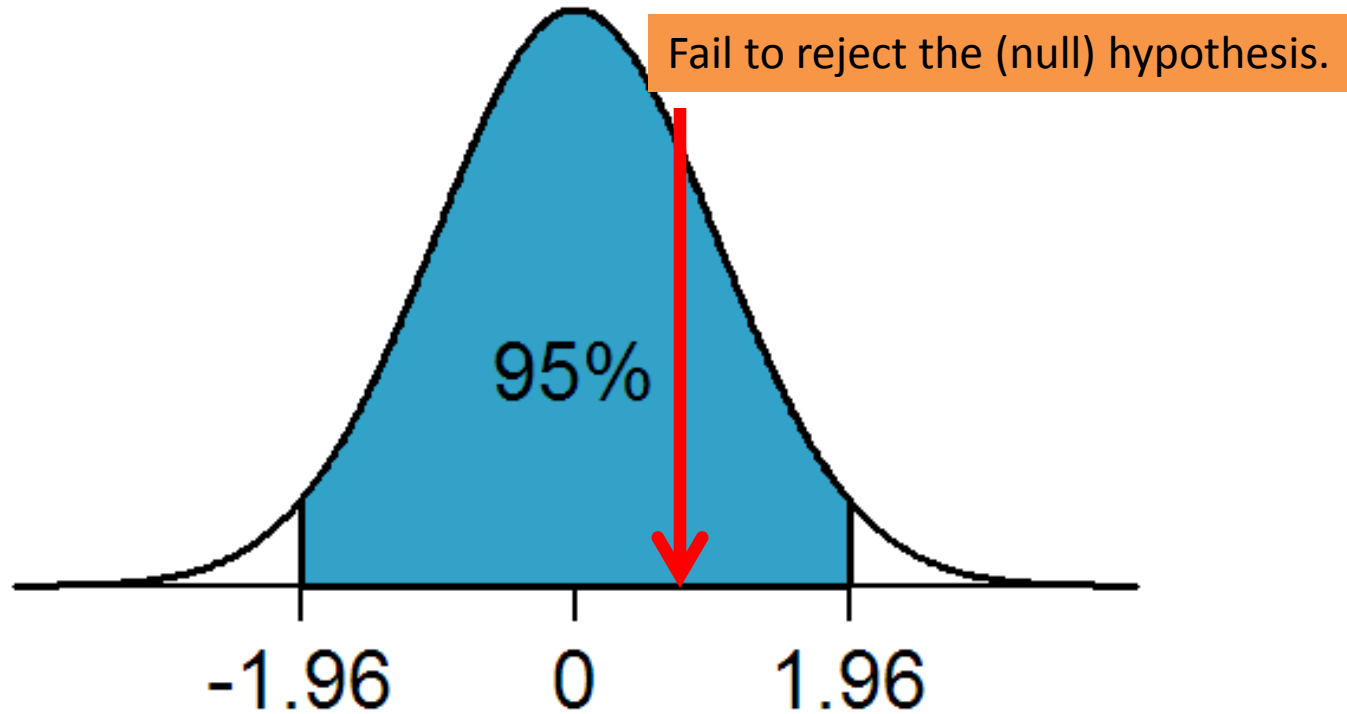
(Step 3) Calculate Statistics From Data



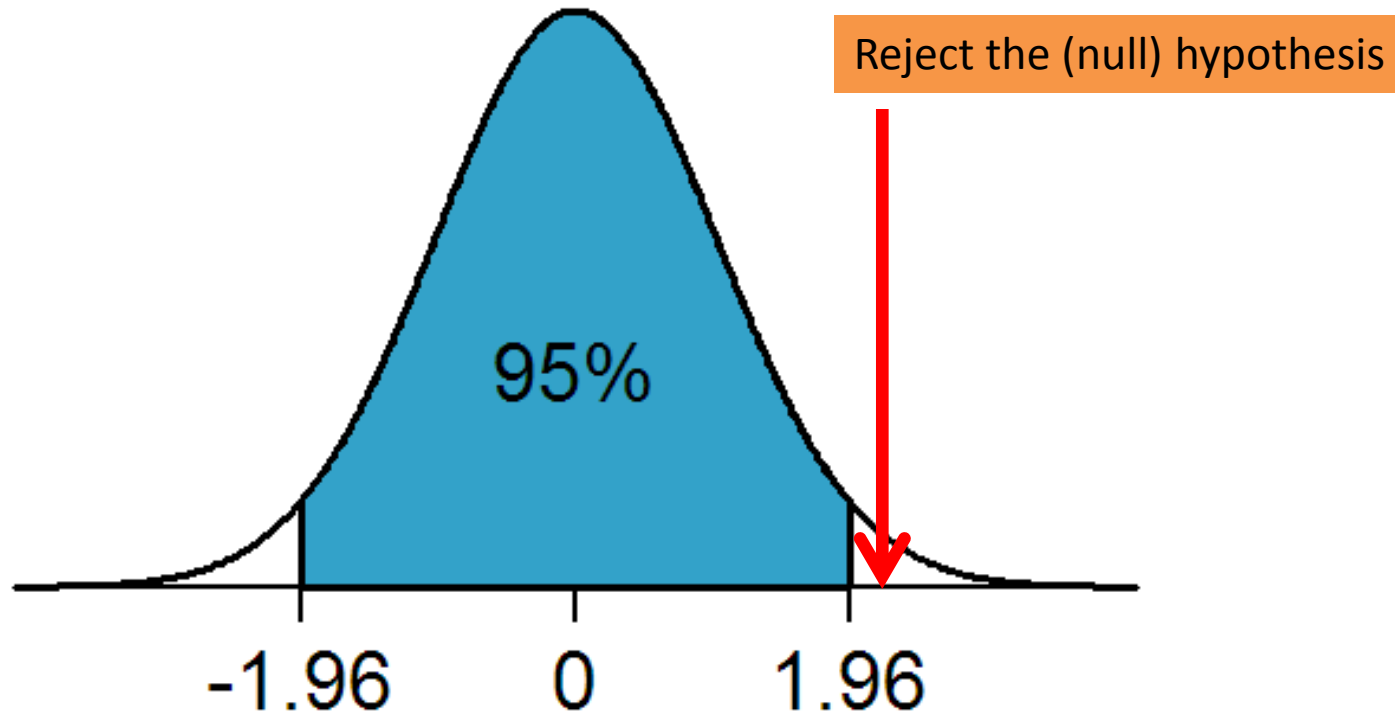
Also called **Test Statistics**.

Test Statistics is usually mean-subtracted. Therefore, we can use mean=0 Normal Distribution to carry on the hypothesis testing.

Hypothesis Testing: (Step 4) Draw Conclusion

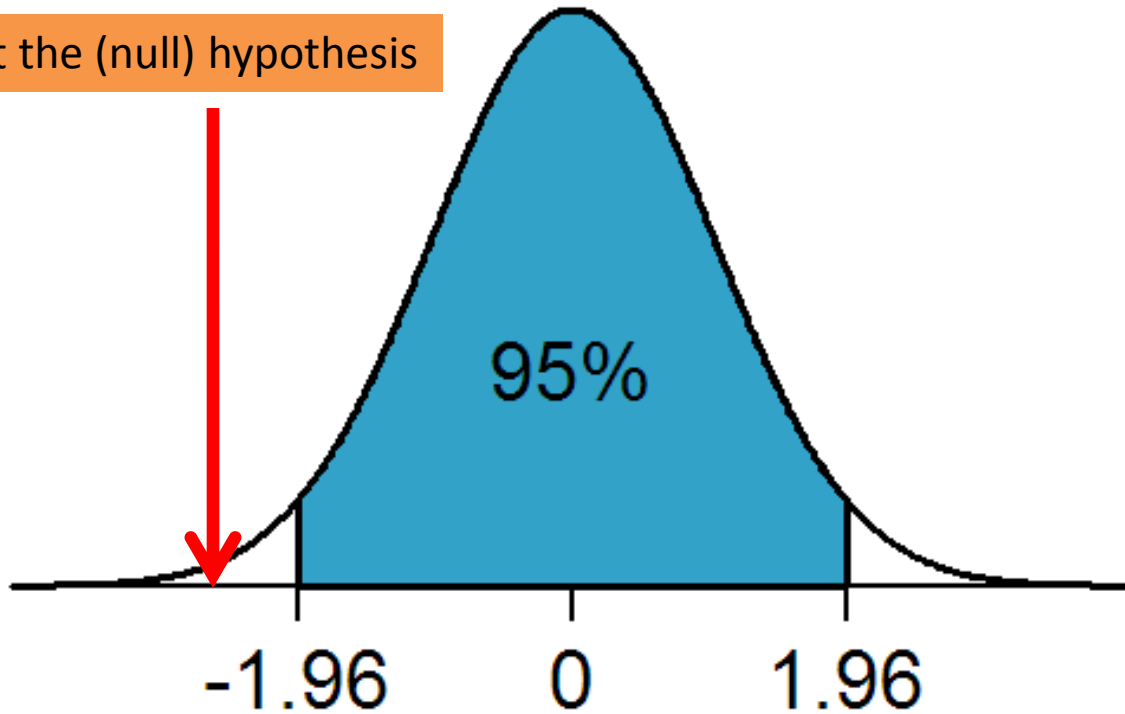


Hypothesis Testing: (Step 4) Draw Conclusion



Hypothesis Testing: (Step 4) Draw Conclusion

Reject the (null) hypothesis



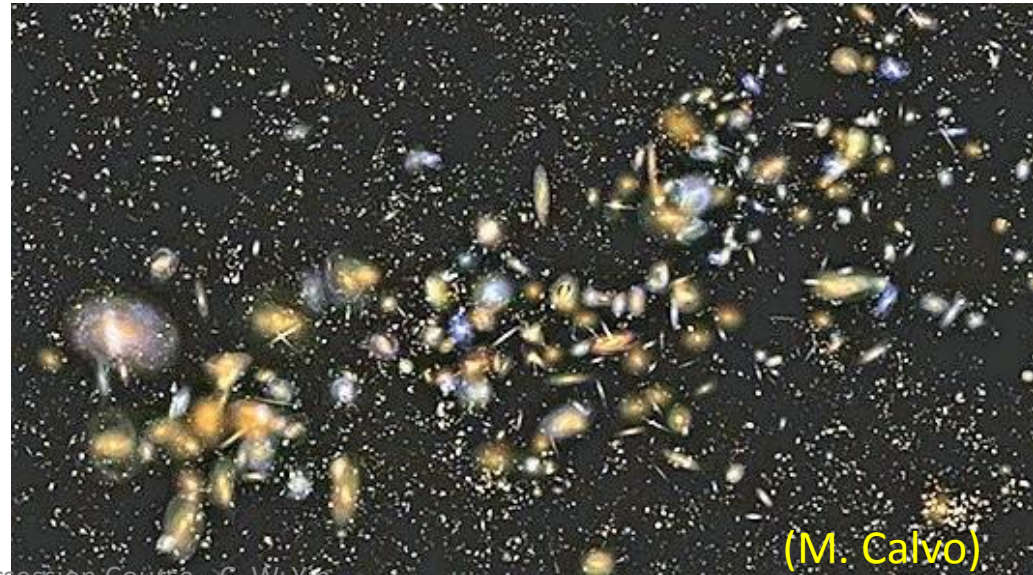
Hypothesis Testing Example

- A galaxy formation theory predicted that the average radius of galaxies in the nearby universe is 35 kpc (1pc = 1parsec = 10^{16} m).
- A random sample of 225 galaxies has a mean radius $\bar{x} = 30$ kpc, and the S.D. of radius = 20 kpc.
- Task: Set up an hypothesis test at 5% significance level.

Null hypothesis: $\mu = 35$ kpc

Alt. hypothesis: $\mu \neq 35$ kpc

Use two-tailed test.



(M. Calvo)

Details:

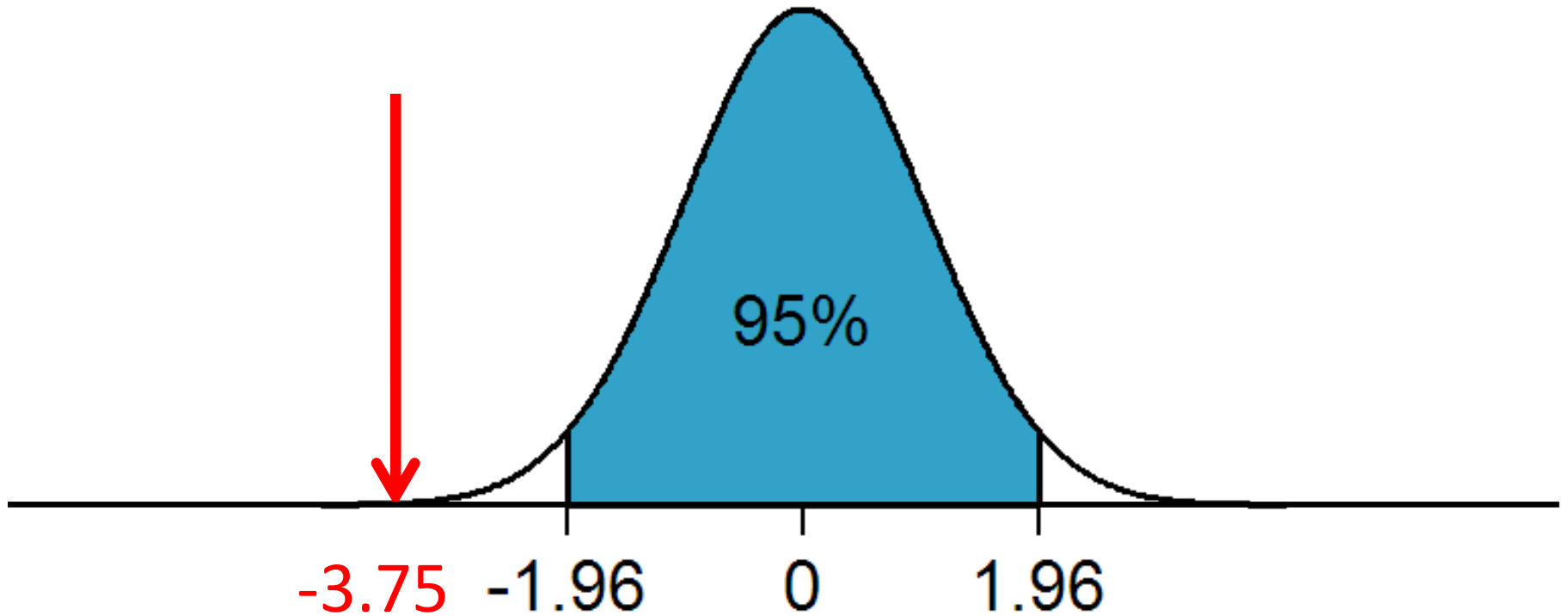
Calculate Sampling Distribution of the mean:

- $\mu_{\bar{x}} = \text{by Central Limit Thm} = \mu(\text{theory}) = 35$
- $\sigma_{\bar{x}} = \text{by Central Limit Thm} = \frac{S.D.(\text{theory})}{\sqrt{n}} \sim \frac{S.D.}{\sqrt{n}} = \frac{20}{15} = \frac{4}{3}$

Calculate Test Statistics from data:

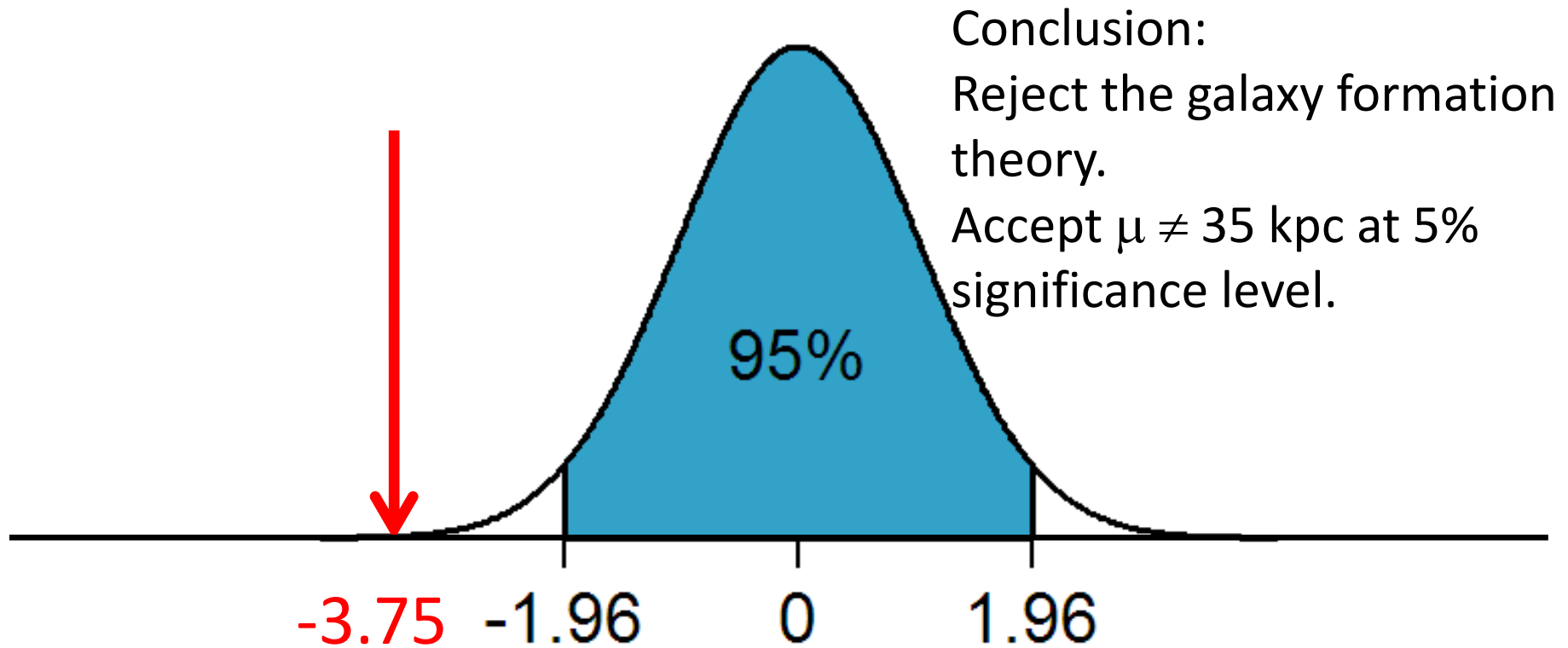
- $Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{30 - 35}{4/3} = -3.75$

$n = 225$ (≥ 30 ; *Almost Normal*)



($\mu = 35$ kpc from theory)

$n = 225$ (≥ 30 ; *Almost Normal*)



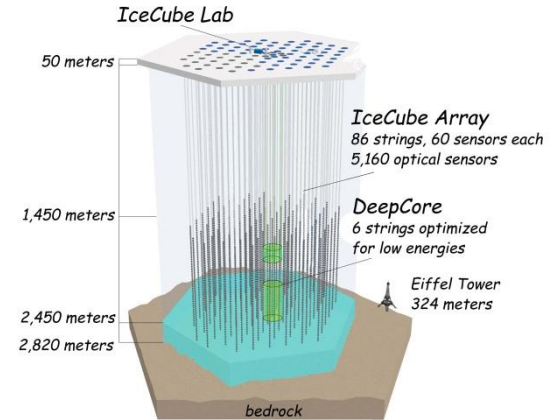
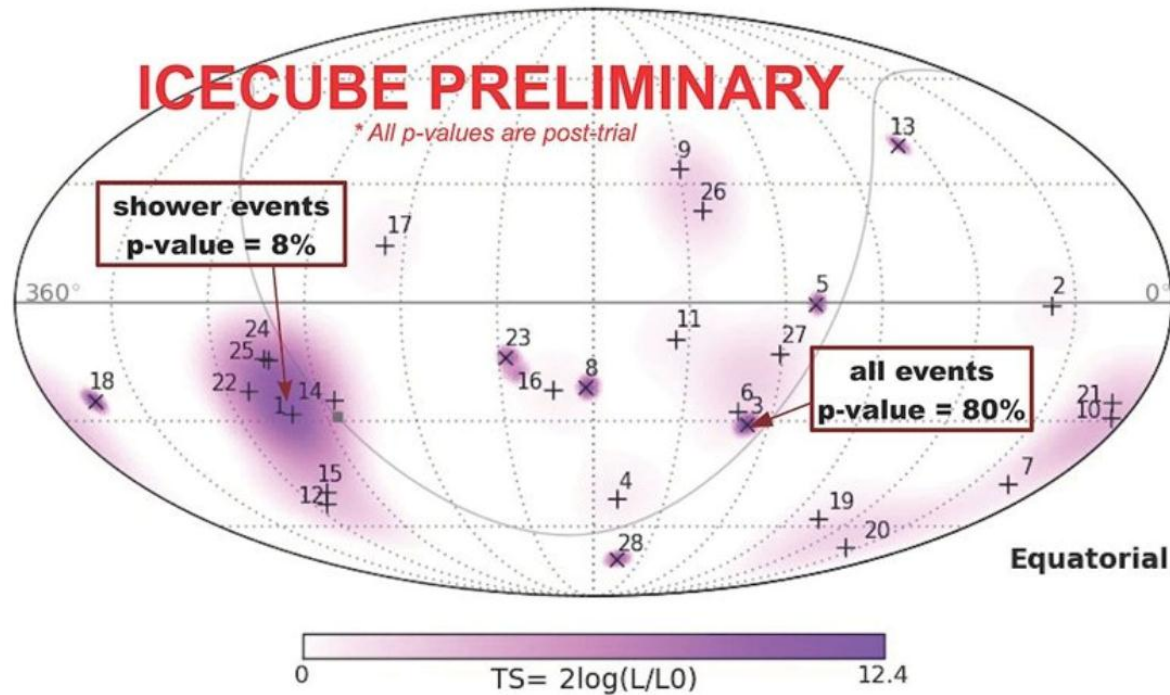
Conclusion:
Reject the galaxy formation theory.
Accept $\mu \neq 35$ kpc at 5% significance level.

($\mu = 35$ kpc from theory)

Meaning of the Confidence Level

- It is the chance we will make **Type I error**.
- Type I error is when we reject Null Hypothesis which is actually true:
 - There is 5% chance we are wrong by rejecting the theory (that average galaxy size being 35 kpc).

Hypothesis Testing Example: 28 Sources in IceCube (2013)



Caption: Sky map in equatorial coordinates of the test statistic (TS) that measures the probability of clustering among the 28 events. The most significant cluster consists of five events—all showers and including the second-highest energy event in the sample—with a final significance of only 8%.

Credit: IceCube Collaboration

Copyright owner is an institution with an existing agreement allowing use by NSF

Submitted by: Francis Halzen

1/14/2014

Image Title: Sky map of 28 extraterrestrial high-energy neutrinos

- **Null Hypothesis:**
 - Sources are uniformly distributed on the sky.
- **Alternative Hypothesis:**
 - Sources are originated from the Milky Way center.

Conditional Probability: Discrete

- Drawing a random card from a stack of 52 playing cards, what is the probability of a card being a Jack given it is a Face card (J, Q, K)?

$$\begin{aligned}P(\text{Jack}|\text{Face}) &= \frac{\text{Number of Jack's}}{\text{Number of Face's}} \\ &= \frac{4}{12} \\ &= \frac{1}{3}\end{aligned}$$



Thinking Like Bayesian

- There will be a bicycle race tomorrow, what is the probability a particular athlete will win?
- Problem: The event only occurs once, we do not have a sample to calculate the probability of a particular athlete winning.
- Solution: Bayesian Statistics.

Bayes' Theorem (1763)

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

Probability of A
before B occurs.

$$\textit{Posterior} = \frac{\textit{Likelihood} * \textit{Prior}}{\textit{Normalization Factor}}$$

Probability of A
after B occurs.



(Thomas Bayes, 1701-1761)

Conditional Probability: Discrete

- Drawing a random card from a stack of 52 playing cards, what is the probability of a card being a Jack given it is a Face card (J, Q, K)?

When I know nothing beforehand:

$$P(\text{Jack}) = \frac{4}{52}$$

If my friend told me it is a Face card (**Added Evidence**):

$$P(\text{Jack}|\text{Face}) = \frac{P(\text{Face}|\text{Jack})P(\text{Jack})}{P(\text{Face})}$$

$$= \frac{1 \cdot \frac{4}{52}}{\frac{12}{52}} = \frac{1}{3}$$

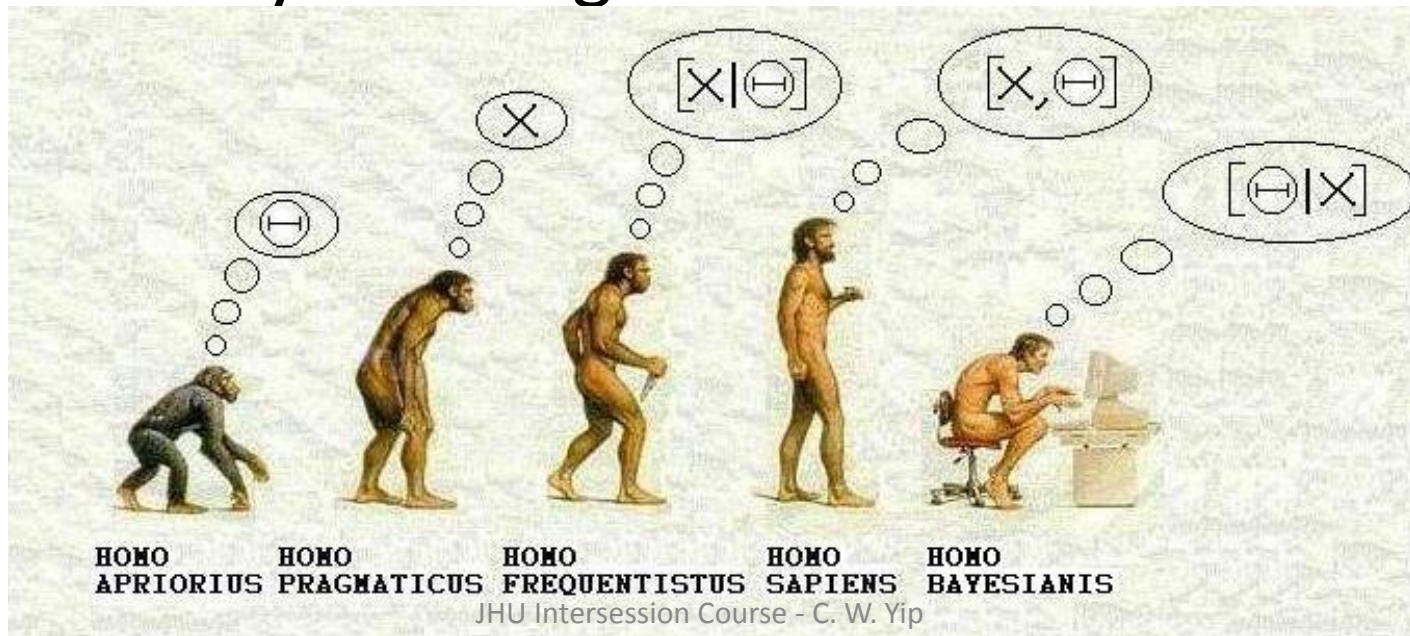


Main Idea behind Bayes' Theorem

*When we add a new piece of evidence,
we change our outlook on the probability of
an event.*

Frequentist vs. Bayesian

- Frequentist: Bayesians assume a prior probability.
- Bayesian: Frequentist cannot assign a probability to a single event.



Parameter Estimation: The Problem

- Estimate parameter from data given model.
- 1 modeled parameter: θ
- Multi modeled parameters: $\vec{\theta} = (\theta_1, \theta_2, \theta_3, \dots, \theta_N)$

Quality of an Estimator: Bias and Variance



unbiased, precise



biased, precise



unbiased, imprecise



biased, imprecise

Usually, the “best estimator”
is somewhere in-between.

Least-Square Fitting (LSQ)

- A popular way to find linear model that best-fit the data.
- A linear model is a model in which there is no square, cube, ..., and higher order power terms in the variables.
- Example: straight lines

Slope and Constant are the *parameters* in the model.

A Parameter Estimation problem.

$$Y = \text{Slope} * X + \text{Constant}$$

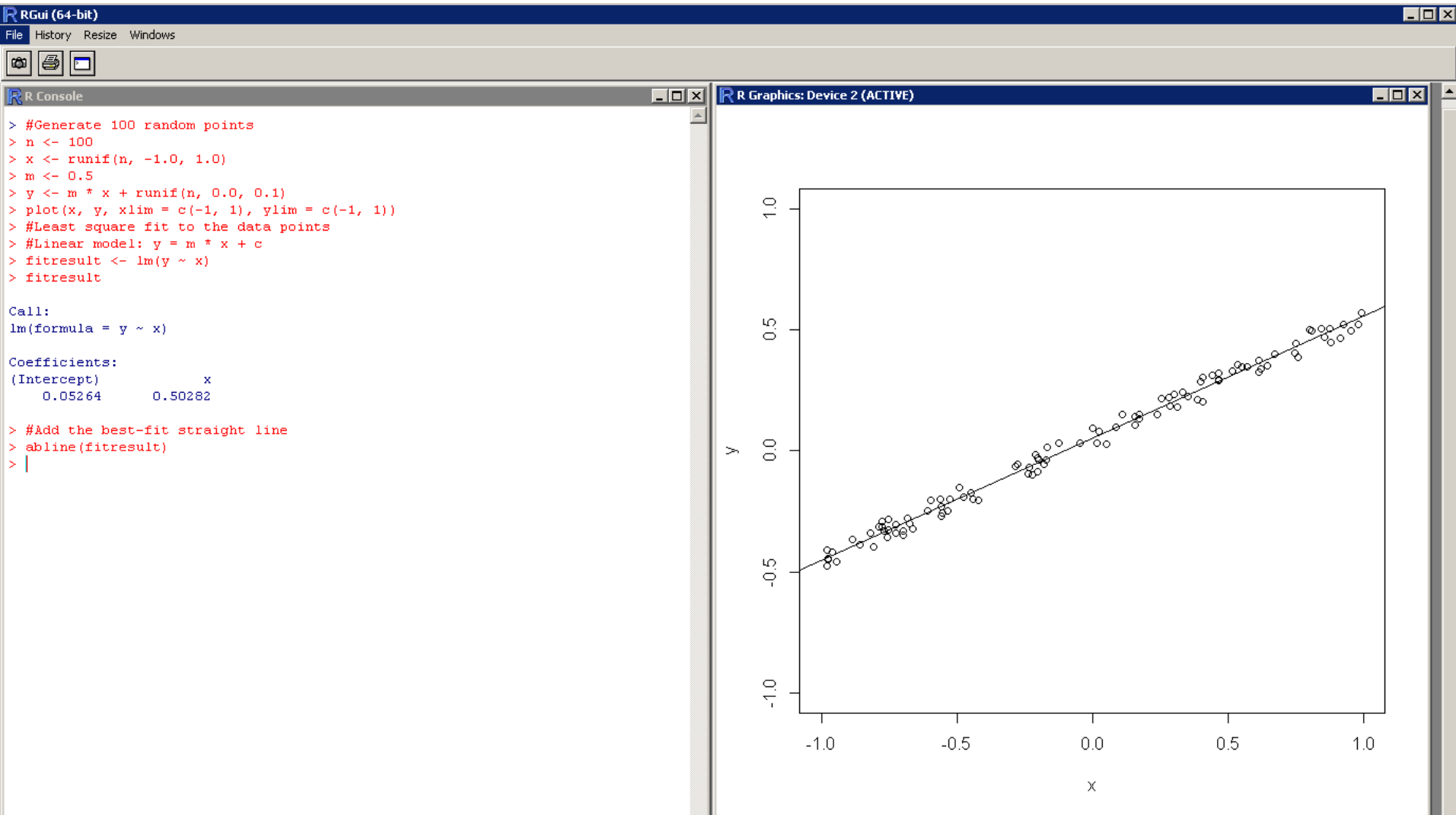
Least-Square Fitting (LSQ): Linear model

$$Y = \text{Slope} * X + \text{Constant}$$

X: Independent Variable, Input Variable, etc.

Y: Dependent Variable, Output Variable, etc.

Example of LSQ Fitting in R

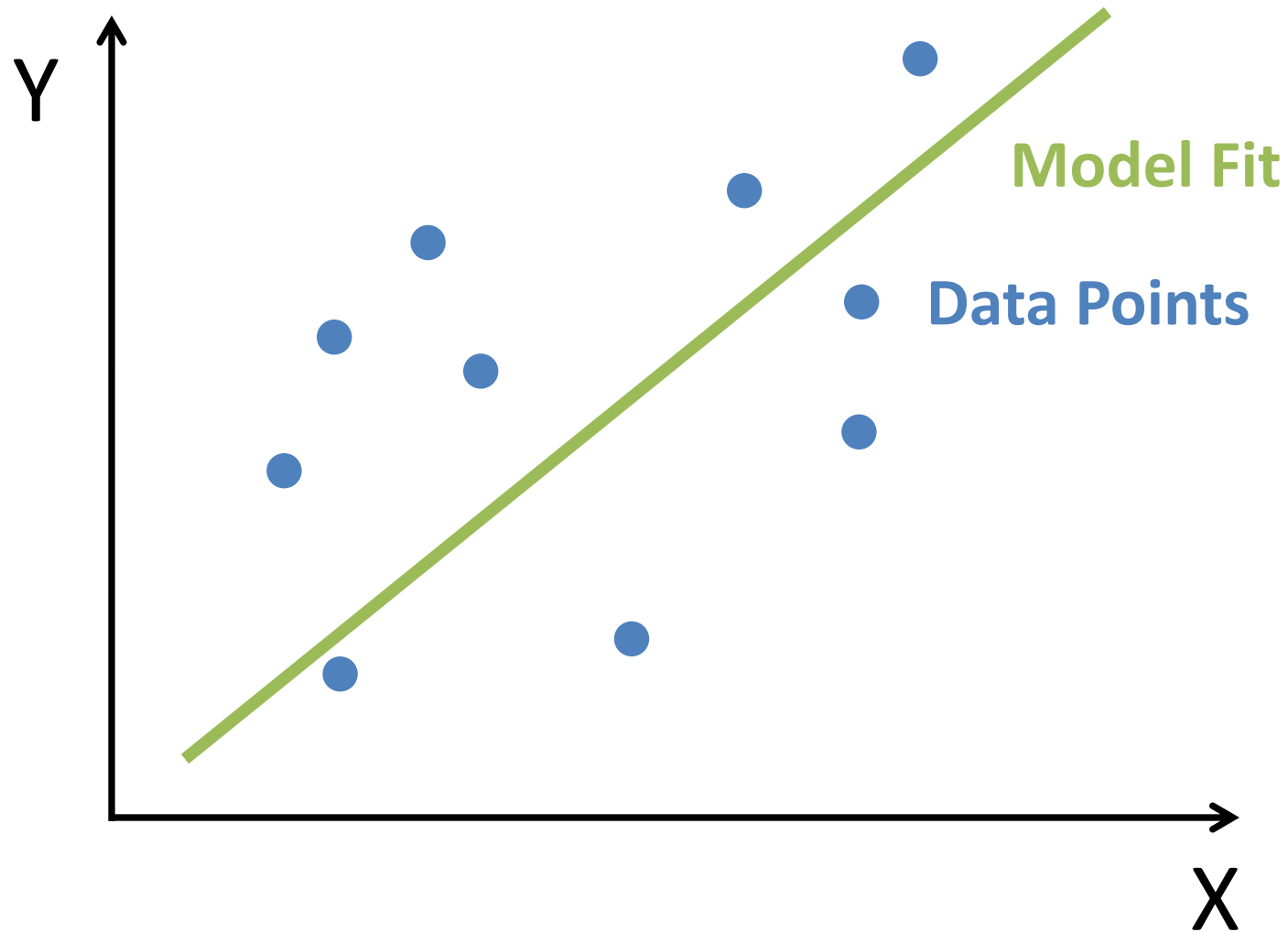


Goodness of Fit

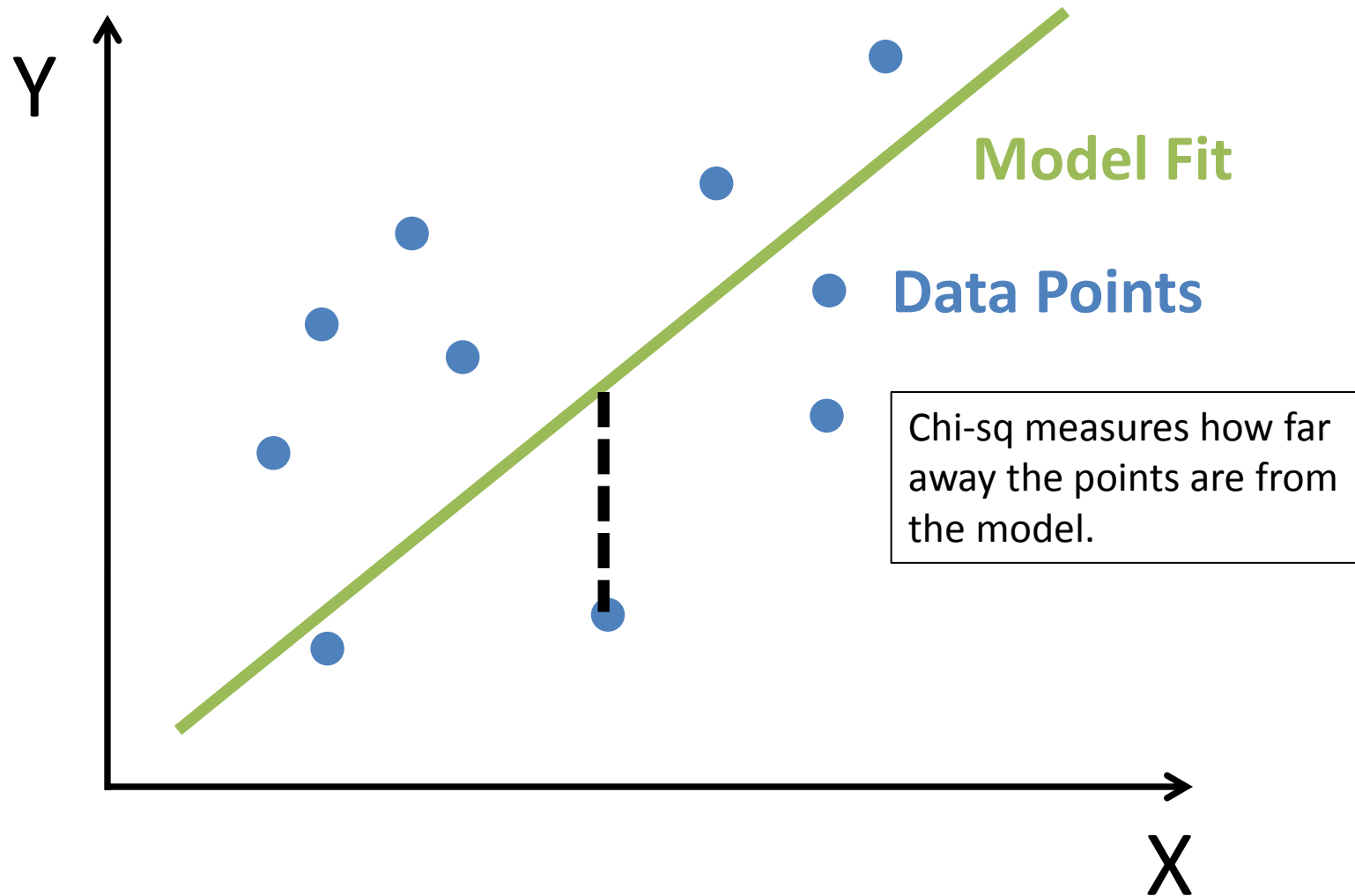
- A goodness of fit measures how well the model fit the data.
- E.g., Chi-sq (the sum of square of the difference over all of the N data points)

$$\chi^2 = \sum_{i=1}^N \frac{(D - M)^2}{\sigma^2}$$

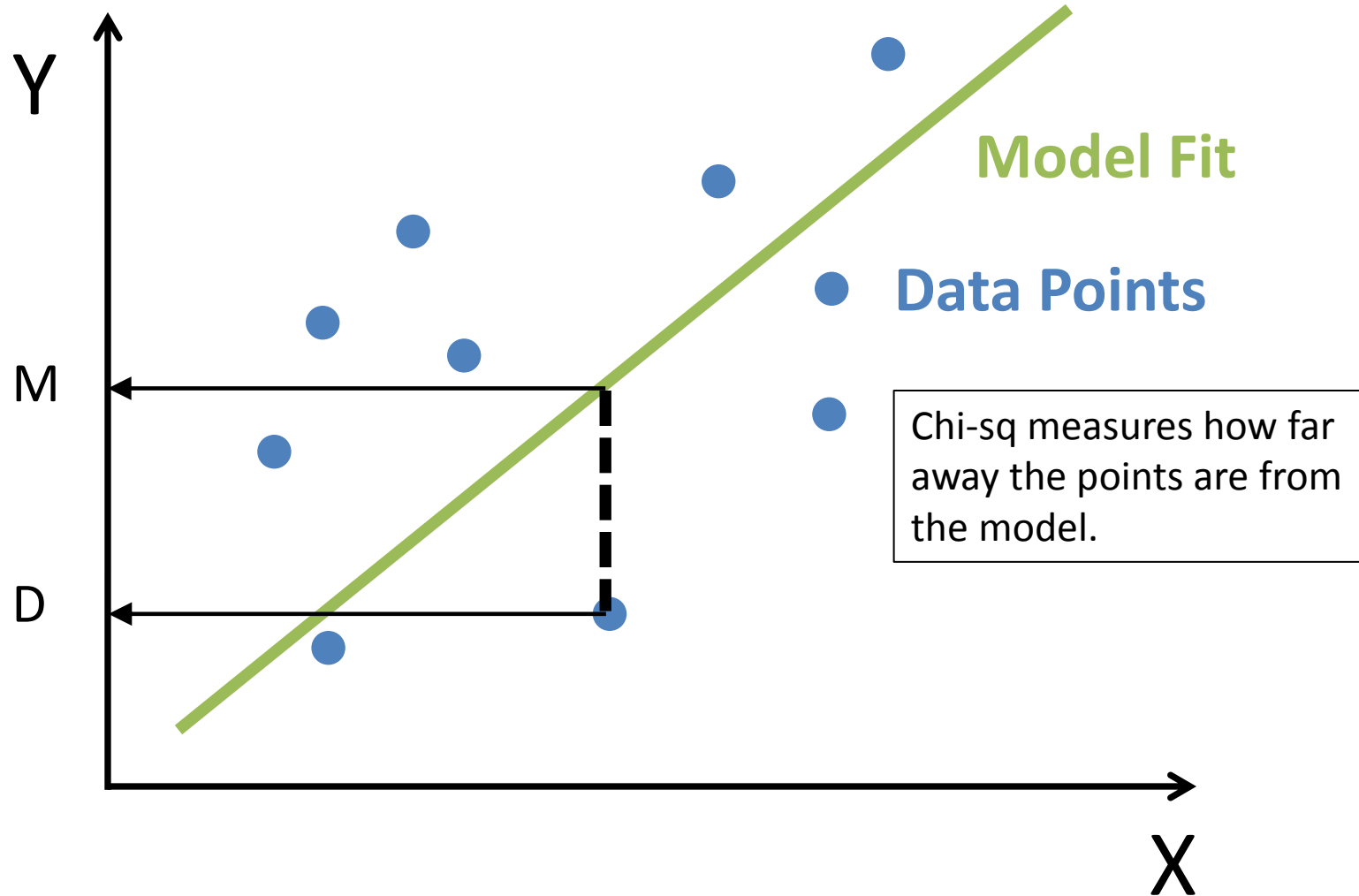
Graphical Meaning of Chi-sq



Graphical Meaning of Chi-sq

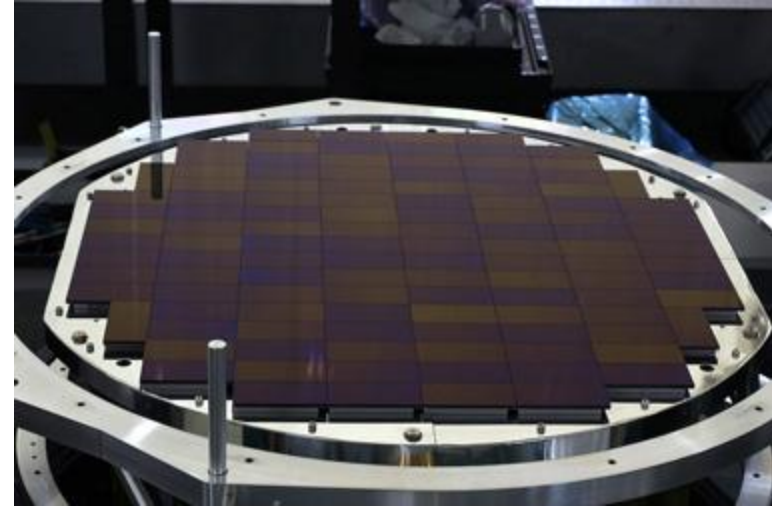


Graphical Meaning of Chi-sq



(Astronomy) Data are Imperfect

- Random Error
 - Photon counts follow Poisson distribution
 - Random error for Poisson = $\sqrt{\textit{photon count}}$
- Systematic Error
 - Bad CCD pixels
 - Cosmic Rays
 - Sky Emissions
 - Etc.



Sky Emission (or Skylines)

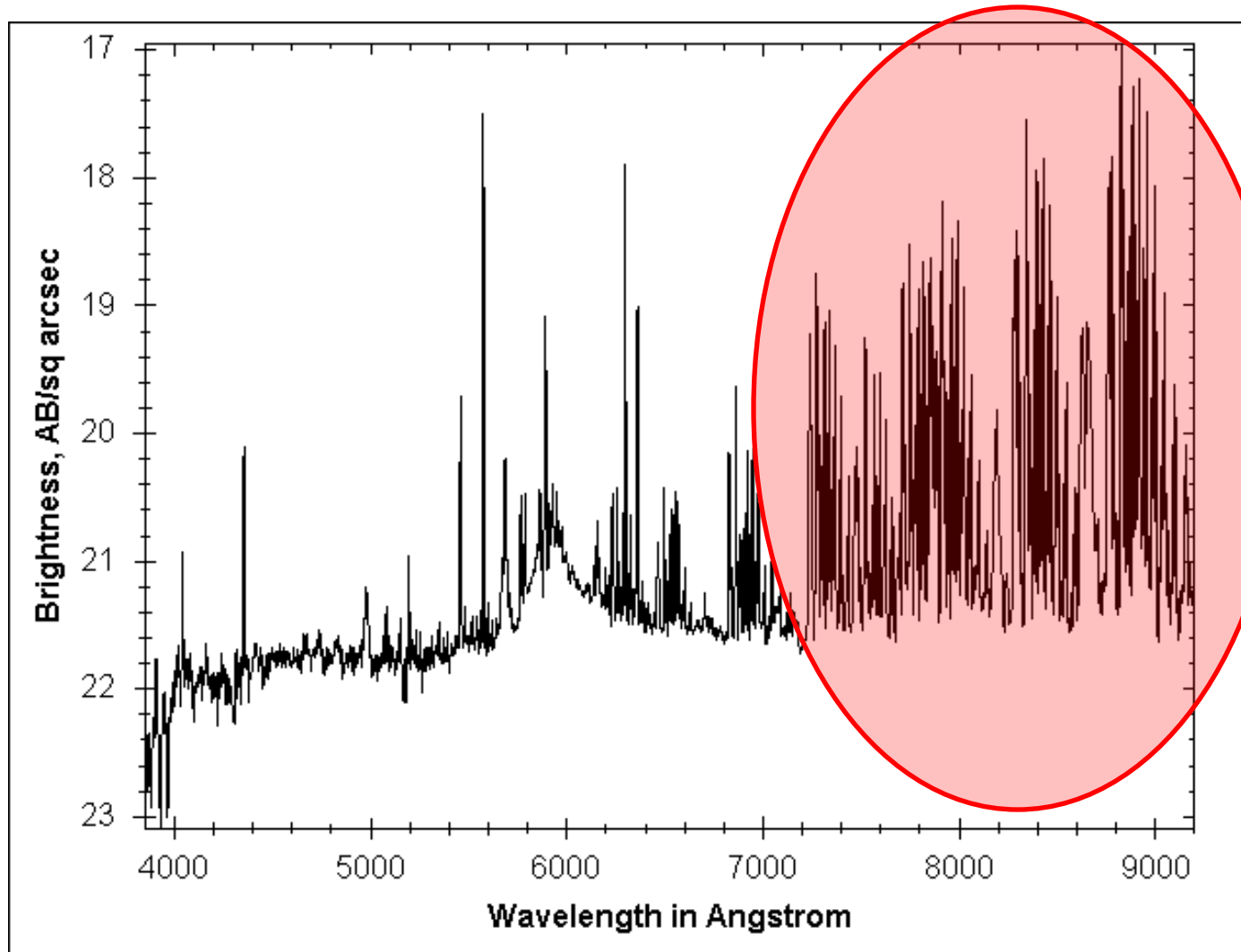


Image Stacking: Central Limit Theorem Revisit

- Fruchter & Hook (2002): Stack images in order to remove cosmic ray (= systematic error in pixel flux)

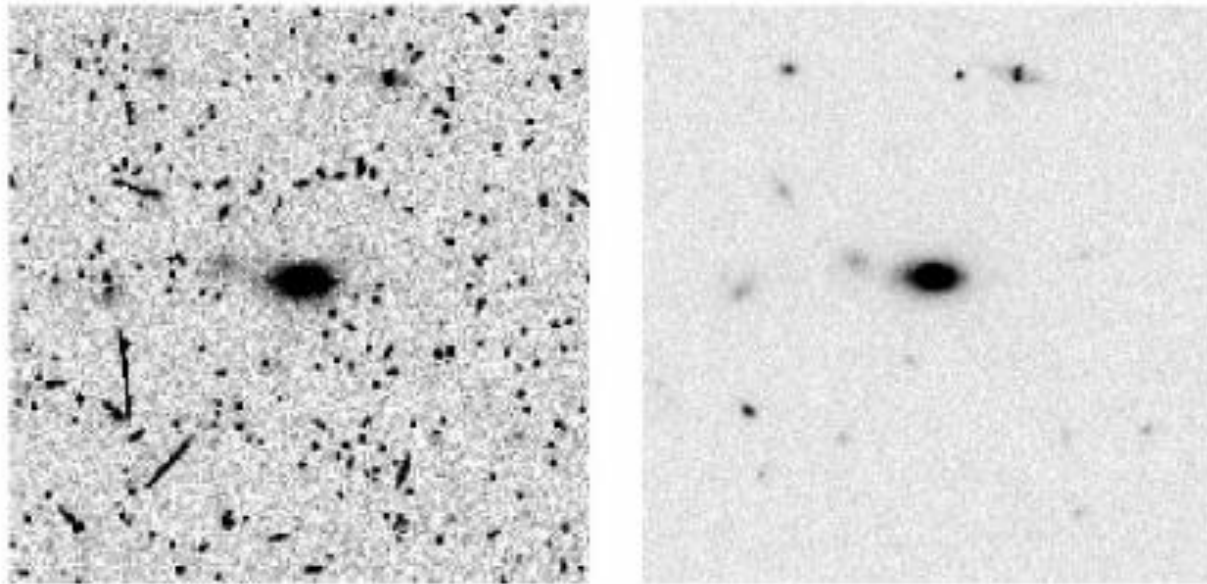
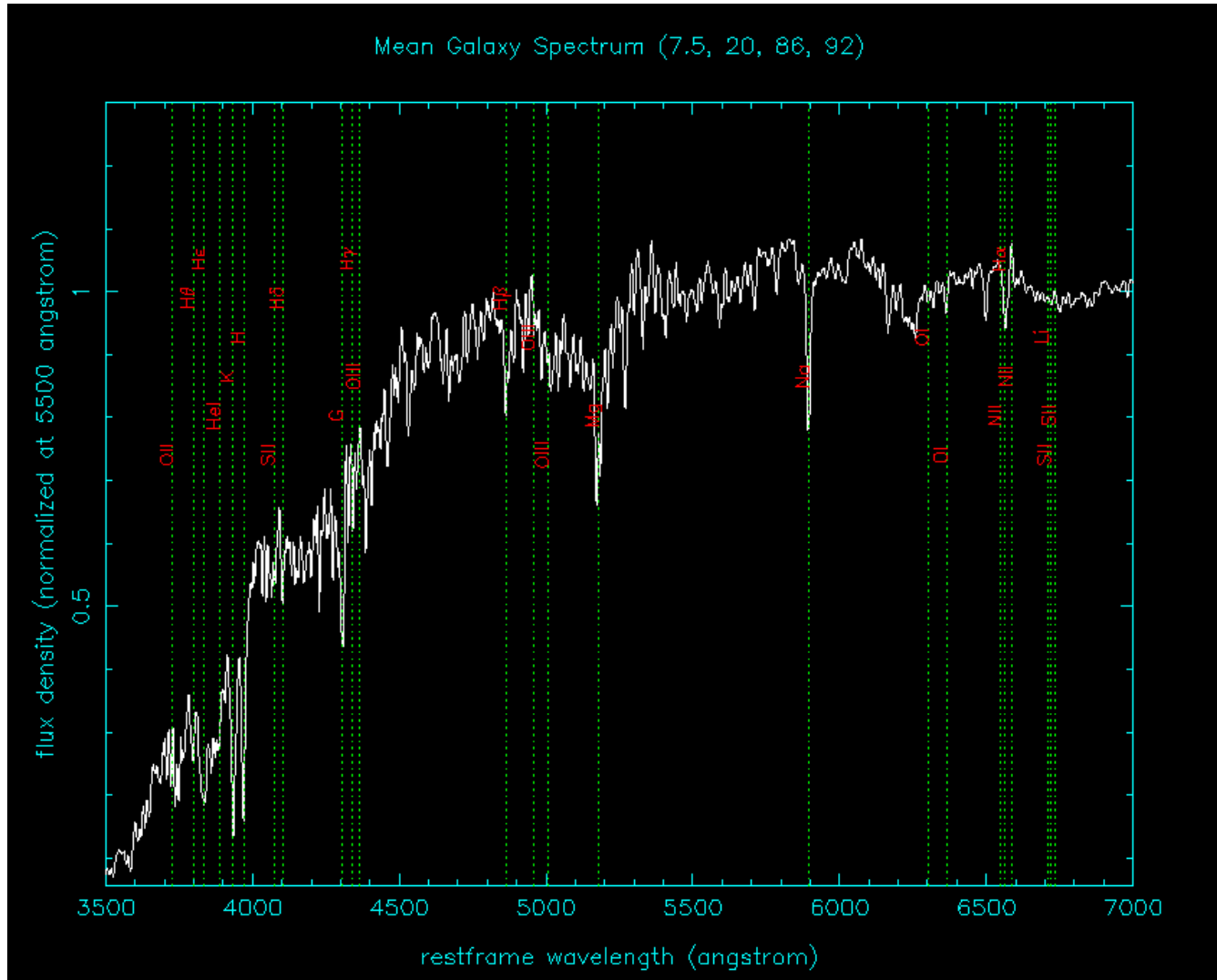
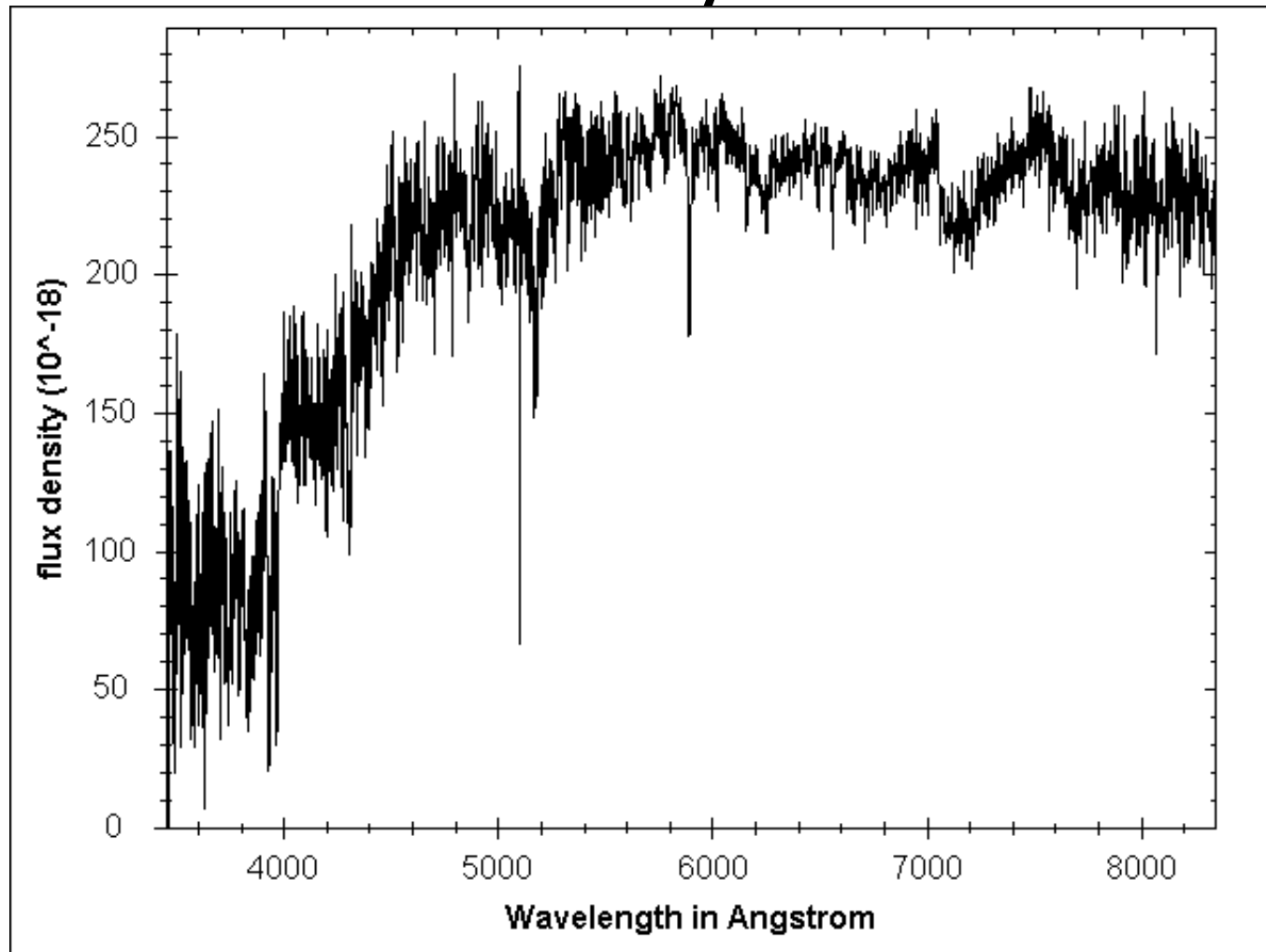


FIG. 3.—*Left:* Region of one of 122400 s archival images taken with the F814W wide near-infrared filter on WFPC2. Numerous cosmic rays are visible. *Right:* Drizzled combination of the 12 images, no two of which shared a dither position.

Similarly, Spectra Stacking



... whereas individual spectrum is
noisy.

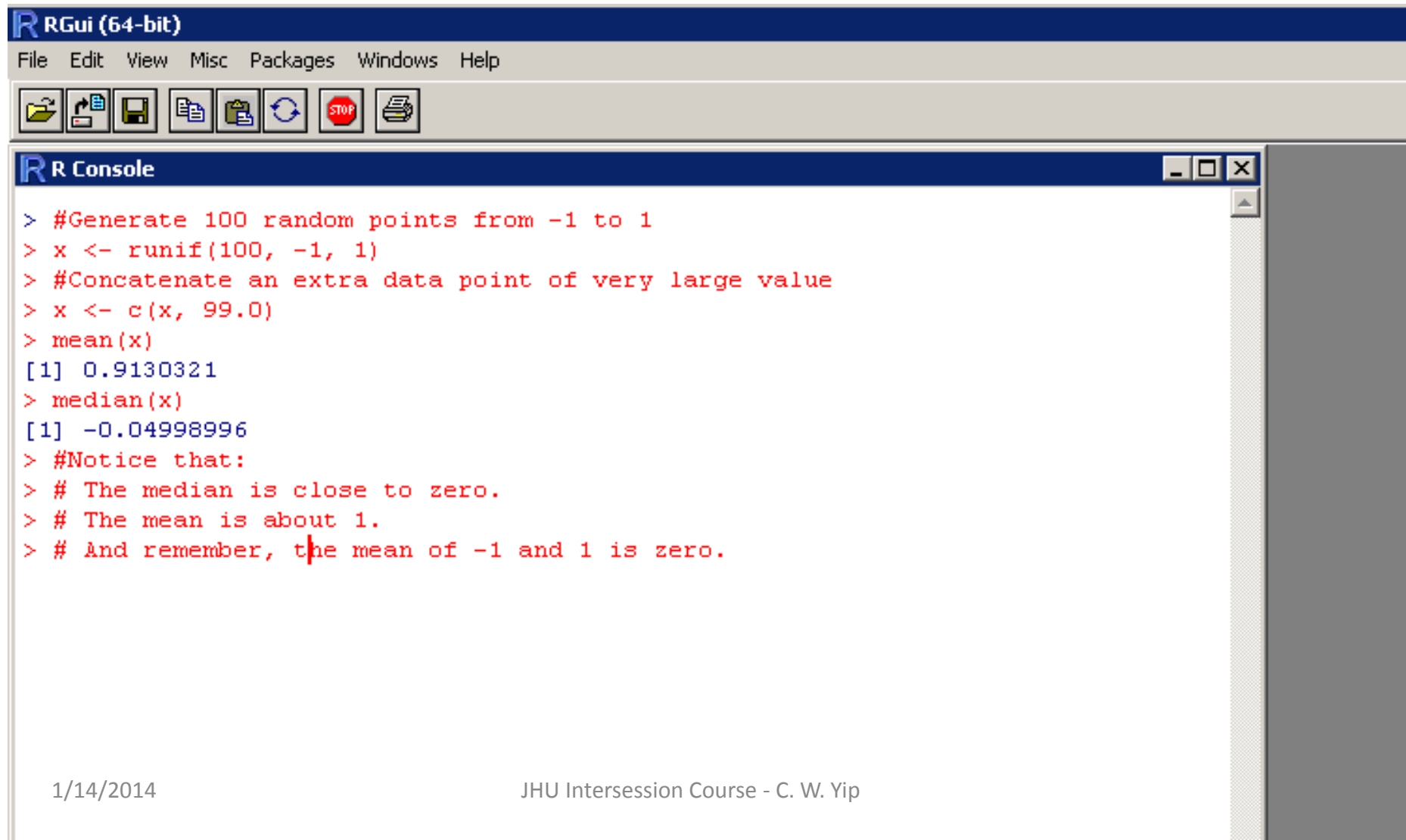


Outliers:

Using Median vs. Mean

- When there are outliers, the median could be more **robust** than the mean as a measure for the average.
- Outliers are difficult to define, because we need to know the average distribution as well (Next Lecture: Unsupervised Machine Learning).
- Subfield of study: *Robust Statistics*.

Median as the Robust Average



The screenshot shows the RGui (64-bit) window with a menu bar (File, Edit, View, Misc, Packages, Windows, Help) and a toolbar. Below the toolbar is the R Console window, which contains the following text:

```
> #Generate 100 random points from -1 to 1
> x <- runif(100, -1, 1)
> #Concatenate an extra data point of very large value
> x <- c(x, 99.0)
> mean(x)
[1] 0.9130321
> median(x)
[1] -0.04998996
> #Notice that:
> # The median is close to zero.
> # The mean is about 1.
> # And remember, the mean of -1 and 1 is zero.
```

Homework

2014 Jan 14 (due Monday noon, Jan 20)

- The data file (saved in the course website as “hubbletable1.csv”) contains the Object Name, Distance (in 10^6 pc = Mpc), and Recession Velocity (in km/s) from Hubble’s 1929 work.
 - 1) Find the Hubble’s Constant by using Least Square Fitting.
 - 2) Plot the Velocity vs. Distance; and the best-fit model.
 - 3) The Hubble’s constant from WMAP survey is determined to be 71 km/s/Mpc. Comment on the comparison between the calculated and the WMAP values.
- Read the article on Bayes’ theorem.
- A CCD records a signal of 100 photons. What is the signal-to-noise ratio (SNR)? If the human eyes can discern features with 100% certainty in an image which has $\text{SNR} \geq 5$ (*), what is the minimum number of photons we need for 100% certainty?

(*) This is a simplified version of the Rose Criterion, 1948.
- Hints:
 - Use `read.csv()` in R to read Comma Separated Values.
 - To extract a column from the data, use `x$column`. For example, `x$Distance_Mpc`.