Stanford

Developing & Teaching Statistics Course for Physics Students

Progress & Lessons Learned

Sanha Cheong
sanha@stanford.edu

APS DSECOP Workshop
June 23, 2022

Context & Motivation

Maybe except some
pure mathematical statisticians

Statistics

- Statistics is, by definition, applied to the real world (*)
- **Statistics is about data**, not just mathematical axioms
- Statistical techniques are highly domain-dependent
  - Physicists, biologists, and economists use very different approaches

Physics

- Physicists are trained with very strong mathematical foundation
- Physicists are experts at quantitatively modeling the real world
  - This is a key skill for many industry jobs
- **We should train physics students with rigorous statistics**

PHYSICS 166/266 @ Stanford

“Statistical Methods in Experimental Physics”

- **Previously, no dedicated statistics foundation course for physicists**
  - Little bit in intro. lab courses
  - Advanced practical tutorial for astro. grad students
  - Otherwise rely on CS & Stats. department
- Primarily for advanced undergrads & junior grad students
- Co-developed and co-taught with Prof. Ariel Schwartzman (sch@slac.stanford.edu)

■ Statistical Methods in Experimental Physics	Winter 2019	TA
■ Statistical Methods in Experimental Physics	Winter 2021	TA
■ Statistical Methods in Experimental Physics	Winter 2022	TA

Philosophy

- **Theoretical foundation**
  - Why/How does *X* work?
  - When does *X* not work? What approximations/assumptions?
  - E.g. “Why is Poisson ~ Gaussian for large *N*?”
    or “Why is this called the χ^2 test statistic?

$$\boxed{\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}}$$

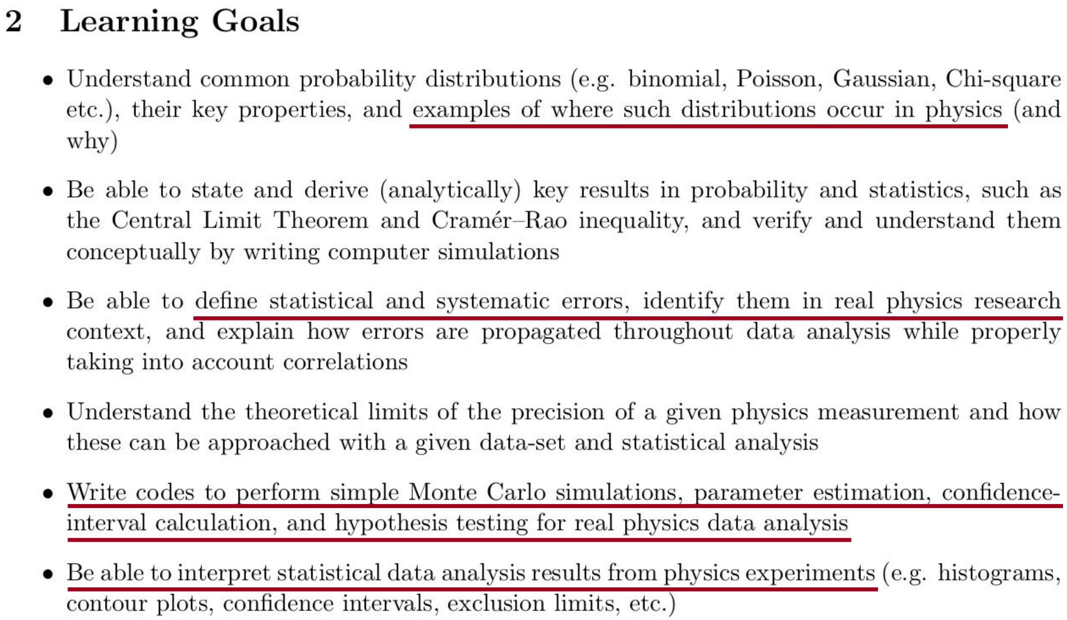
- **Computational practice**
  - Practice empowers students, gives them confidence, etc.
  - Necessary for research, job market, etc.
- **Primary goal: training for future experimental physicists**
  - With strong foundation, specific advanced techniques can be learned easily
    (in specific research groups, with special topics courses, etc.)

Design Consideration

- Prerequisites:
  - Most students come with some coding literacy
  - Solid foundation on multi-variable calculus and linear algebra

    (exposure to Fourier transform is good, but not strictly necessary)
  - At least two years of undergrad physics
    - Conceptual understanding of QM
    - E.g. there is fundamental randomness in nature, there are discrete states
- 10 weeks total (quarter system)
  - **3 class meetings / week, 1 for Jupyter Notebook sessions** (Google Colab)
- What this course is NOT:
  - Overview of all methods in current research
  - A course on machine learning

Syllabus



2 Learning Goals

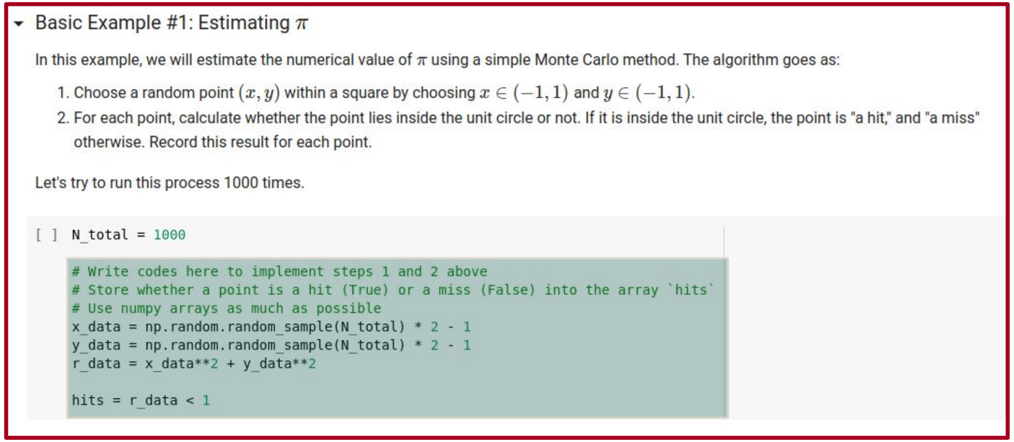
- Understand common probability distributions (e.g. binomial, Poisson, Gaussian, Chi-square etc.), their key properties, and examples of where such distributions occur in physics (and why)
- Be able to state and derive (analytically) key results in probability and statistics, such as the Central Limit Theorem and Cramér–Rao inequality, and verify and understand them conceptually by writing computer simulations
- Be able to define statistical and systematic errors, identify them in real physics research context, and explain how errors are propagated throughout data analysis while properly taking into account correlations
- Understand the theoretical limits of the precision of a given physics measurement and how these can be approached with a given data-set and statistical analysis
- Write codes to perform simple Monte Carlo simulations, parameter estimation, confidence-interval calculation, and hypothesis testing for real physics data analysis
- Be able to interpret statistical data analysis results from physics experiments (e.g. histograms, contour plots, confidence intervals, exclusion limits, etc.)

QR Code for Syllabus

Examples of Course Materials

Jupyter Notebooks, Homework Problems / Solutions

Introduction to Monte Carlo Methods (1/2)



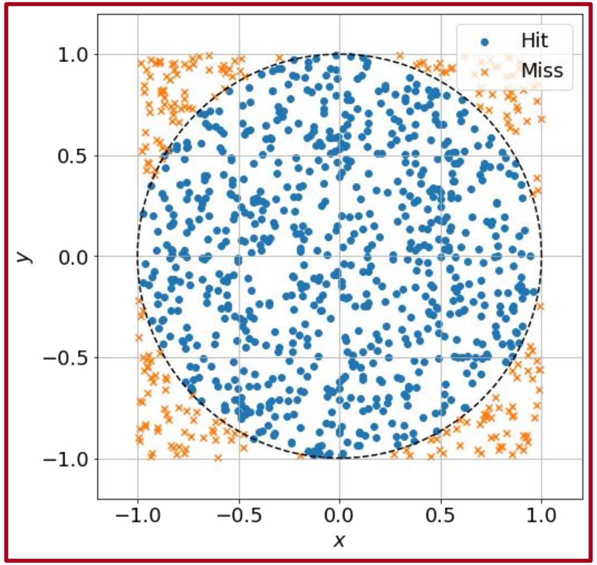
Basic Example #1: Estimating π

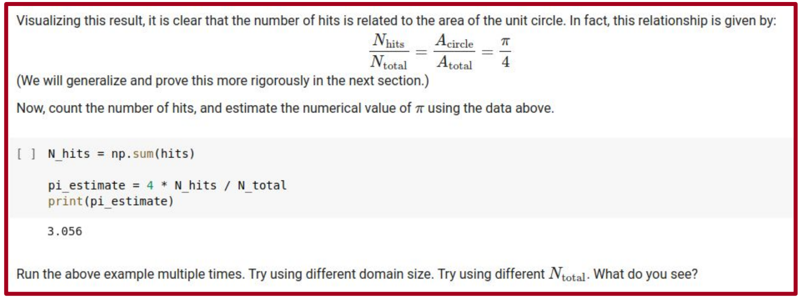
In this example, we will estimate the numerical value of π using a simple Monte Carlo method. The algorithm goes as:

1. Choose a random point (x, y) within a square by choosing $x \in (-1, 1)$ and $y \in (-1, 1)$.
2. For each point, calculate whether the point lies inside the unit circle or not. If it is inside the unit circle, the point is "a hit," and "a miss" otherwise. Record this result for each point.

Let's try to run this process 1000 times.

```
[ ] N_total = 1000

    # Write codes here to implement steps 1 and 2 above
    # Store whether a point is a hit (True) or a miss (False) into the array `hits`
    # Use numpy arrays as much as possible
    x_data = np.random.random_sample(N_total) * 2 - 1
    y_data = np.random.random_sample(N_total) * 2 - 1
    r_data = x_data**2 + y_data**2

    hits = r_data < 1
```





Visualizing this result, it is clear that the number of hits is related to the area of the unit circle. In fact, this relationship is given by:

$$\frac{N_{\text{hits}}}{N_{\text{total}}} = \frac{A_{\text{circle}}}{A_{\text{total}}} = \frac{\pi}{4}$$

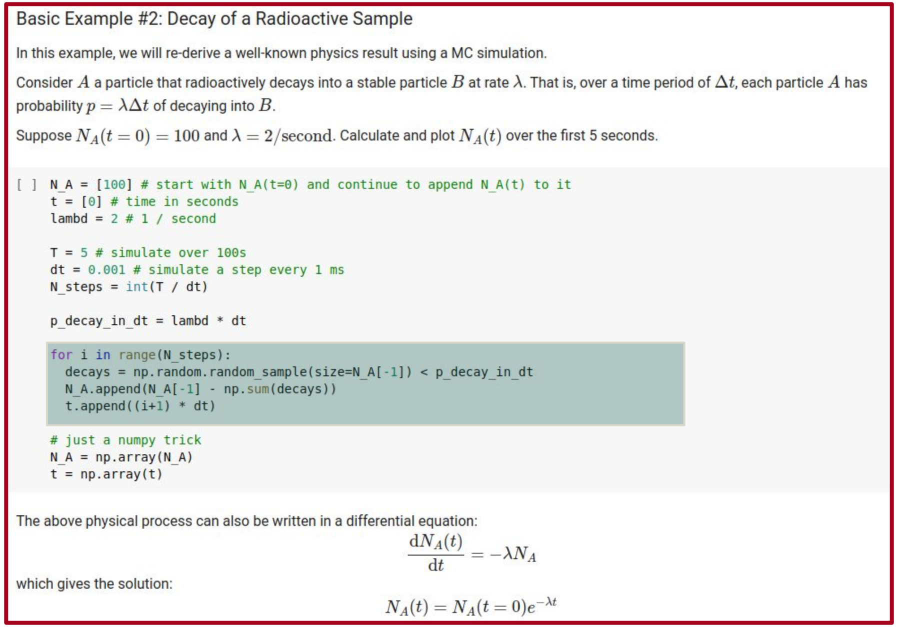
(We will generalize and prove this more rigorously in the next section.)

Now, count the number of hits, and estimate the numerical value of π using the data above.

```
[ ] N_hits = np.sum(hits)

    pi_estimate = 4 * N_hits / N_total
    print(pi_estimate)

    3.056
```

Run the above example multiple times. Try using different domain size. Try using different N_{total}. What do you see?

Introduction to Monte Carlo Methods (2/2)



Basic Example #2: Decay of a Radioactive Sample

In this example, we will re-derive a well-known physics result using a MC simulation.

Consider A a particle that radioactively decays into a stable particle B at rate λ. That is, over a time period of Δt, each particle A has probability $p = \lambda \Delta t$ of decaying into B.

Suppose $N_A(t=0) = 100$ and $\lambda = 2/\mathrm{second}$. Calculate and plot $N_A(t)$ over the first 5 seconds.

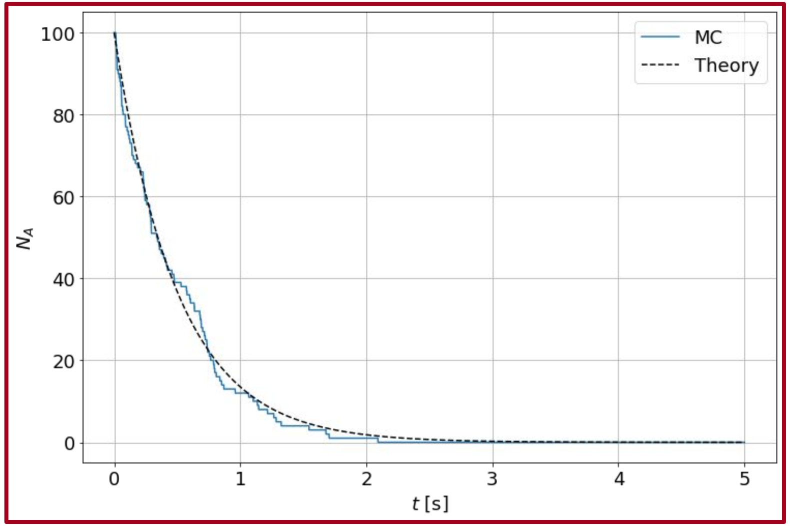
```
[ ]  N_A = [100] # start with N_A(t=0) and continue to append N_A(t) to it
     t = [0] # time in seconds
     lambd = 2 # 1 / second

     T = 5 # simulate over 100s
     dt = 0.001 # simulate a step every 1 ms
     N_steps = int(T / dt)

     p_decay_in_dt = lambd * dt

     for i in range(N_steps):
       decays = np.random.random_sample(size=N_A[-1]) < p_decay_in_dt
       N_A.append(N_A[-1] - np.sum(decays))
       t.append((i+1) * dt)

     # just a numpy trick
     N_A = np.array(N_A)
     t = np.array(t)
```

The above physical process can also be written in a differential equation:

$$\frac{\mathrm{d}N_A(t)}{\mathrm{d}t} = -\lambda N_A$$

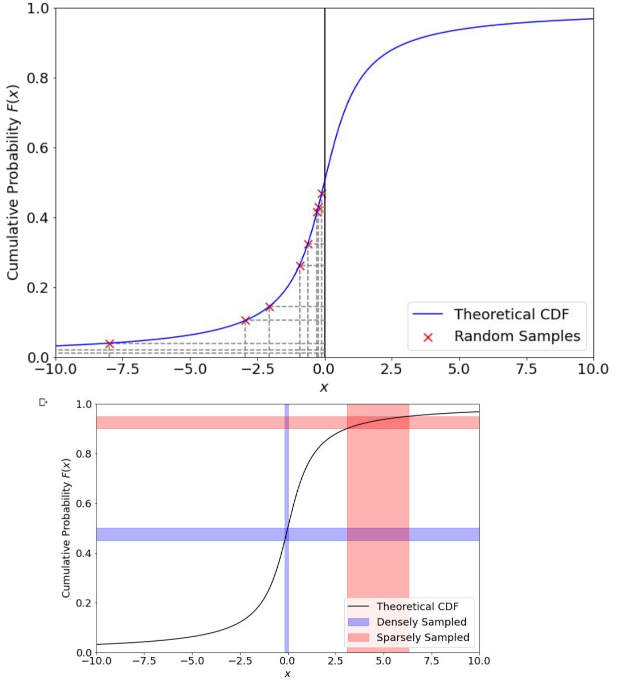
which gives the solution:

$$N_A(t) = N_A(t=0)e^{-\lambda t}$$



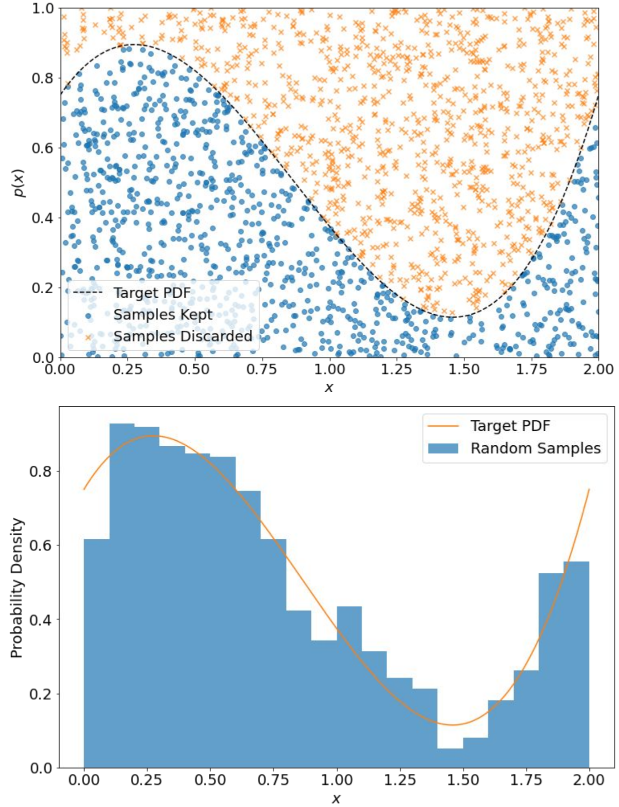
Sampling from Known Distributions

Inverse Transform Sampling



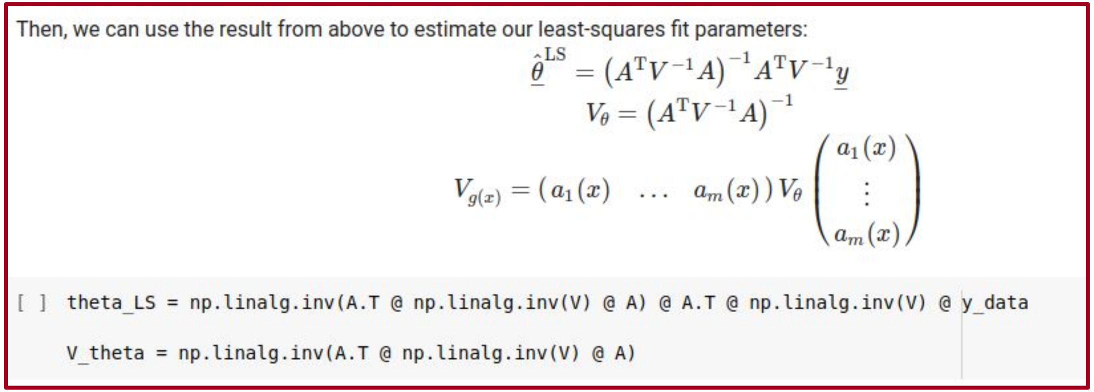
If we follow the above process, we can visualize that a small interval δu on the vertical axis will be stretched onto a larger interval on the horizontal axis if δu lies near the range where $F_X(x)$ is very flat---i.e. $F_X^{-1}(u)$ is very steep. Then, the x-region corresponding to this interval has low density. Vice versa, near whre $F_X(x)$ is steep, $F_X^{-1}(x)$ is flat, so the vertical interval δu gets squeezed onto a narrow horizontal interval, leading to high density.

This conversion factor is precisely related to the derivative of the inverse CDF. This is why the inverse-transform works. (It follows the exactly same line of logic as the equation for the transformation of variables.)



Rejection Sampling

Model Fitting & Combining Measurements



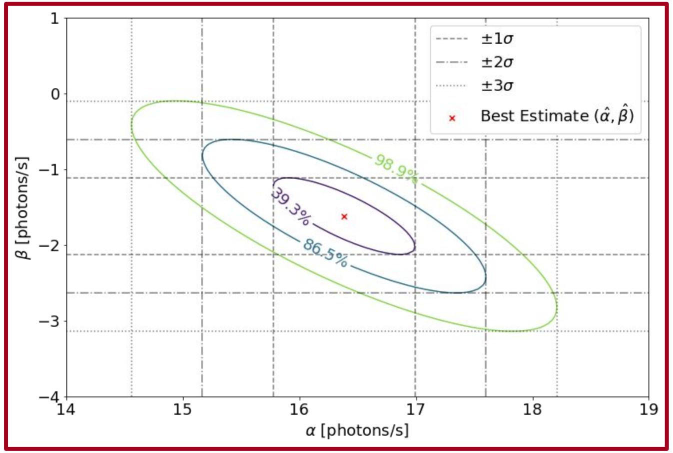
Then, we can use the result from above to estimate our least-squares fit parameters:

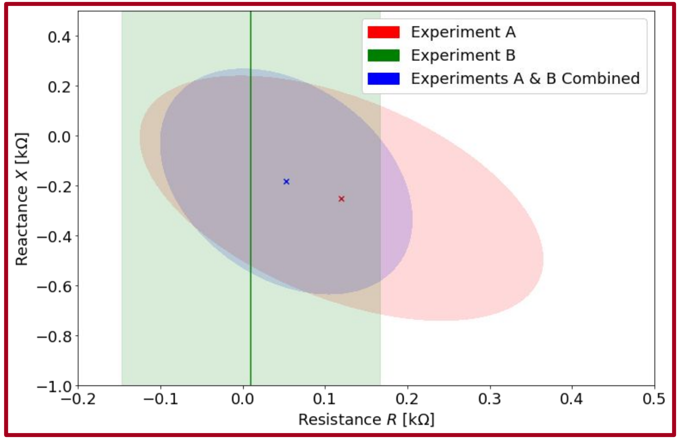
$$\underline{\hat{\theta}}^{\mathrm{LS}} = \left(A^{\mathrm{T}}V^{-1}A\right)^{-1}A^{\mathrm{T}}V^{-1}\underline{y}$$

$$V_{\theta} = \left(A^{\mathrm{T}}V^{-1}A\right)^{-1}$$

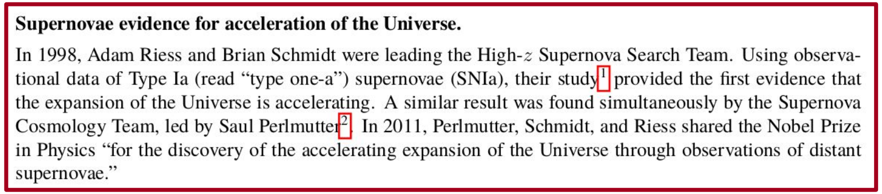
$$V_{g(x)} = \begin{pmatrix} a_1(x) & \ldots & a_m(x) \end{pmatrix} V_{\theta} \begin{pmatrix} a_1(x) \\ \vdots \\ a_m(x) \end{pmatrix}$$

```
[ ]  theta_LS = np.linalg.inv(A.T @ np.linalg.inv(V) @ A) @ A.T @ np.linalg.inv(V) @ y_data

     V_theta = np.linalg.inv(A.T @ np.linalg.inv(V) @ A)
```



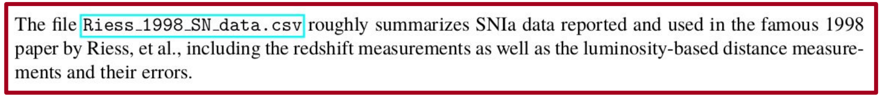


Replicate Nobel Prize Winning Discovery: SNIa Evidence for Expanding Universe (1/2)

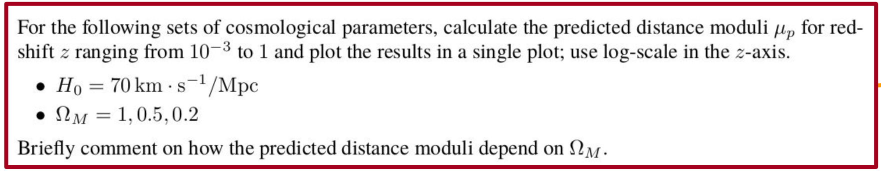


Supernovae evidence for acceleration of the Universe.

In 1998, Adam Riess and Brian Schmidt were leading the High-z Supernova Search Team. Using observational data of Type Ia (read "type one-a") supernovae (SNIa), their study[1] provided the first evidence that the expansion of the Universe is accelerating. A similar result was found simultaneously by the Supernova Cosmology Team, led by Saul Perlmutter[2]. In 2011, Perlmutter, Schmidt, and Riess shared the Nobel Prize in Physics "for the discovery of the accelerating expansion of the Universe through observations of distant supernovae."



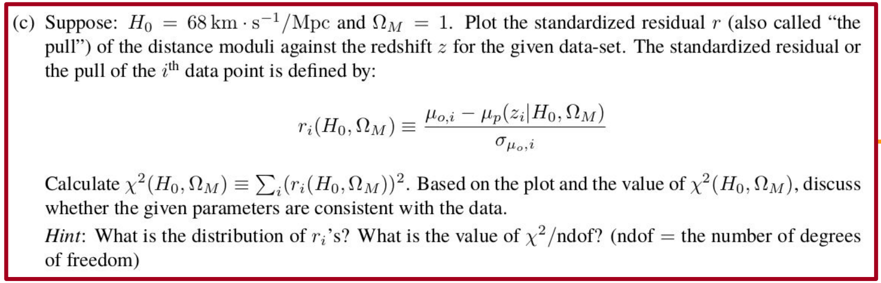
The file `Riess_1998_SN_data.csv` roughly summarizes SNIa data reported and used in the famous 1998 paper by Riess, et al., including the redshift measurements as well as the luminosity-based distance measurements and their errors.



For the following sets of cosmological parameters, calculate the predicted distance moduli μ_p for redshift z ranging from 10^{-3} to 1 and plot the results in a single plot; use log-scale in the z-axis.

- $H_0 = 70\,\mathrm{km}\cdot\mathrm{s}^{-1}/\mathrm{Mpc}$
- $\Omega_M = 1, 0.5, 0.2$

Briefly comment on how the predicted distance moduli depend on Ω_M.

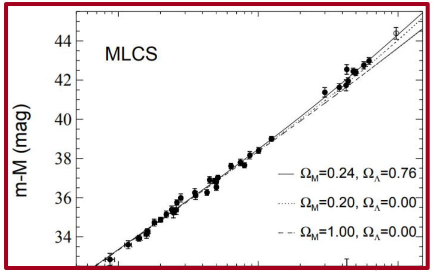


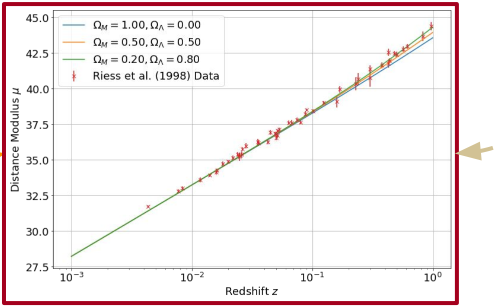
(c) Suppose: $H_0 = 68\,\mathrm{km}\cdot\mathrm{s}^{-1}/\mathrm{Mpc}$ and $\Omega_M = 1$. Plot the standardized residual r (also called "the pull") of the distance moduli against the redshift z for the given data-set. The standardized residual or the pull of the i^{th} data point is defined by:

$$r_i(H_0, \Omega_M) \equiv \frac{\mu_{o,i} - \mu_p(z_i|H_0, \Omega_M)}{\sigma_{\mu_o,i}}$$

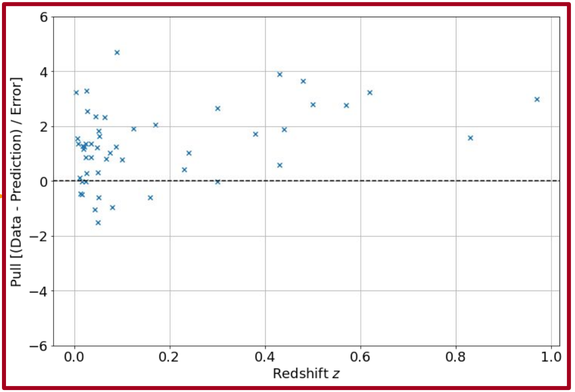
Calculate $\chi^2(H_0, \Omega_M) \equiv \sum_i (r_i(H_0, \Omega_M))^2$. Based on the plot and the value of $\chi^2(H_0, \Omega_M)$, discuss whether the given parameters are consistent with the data.

Hint: What is the distribution of r_i's? What is the value of χ^2/ndof? (ndof = the number of degrees of freedom)



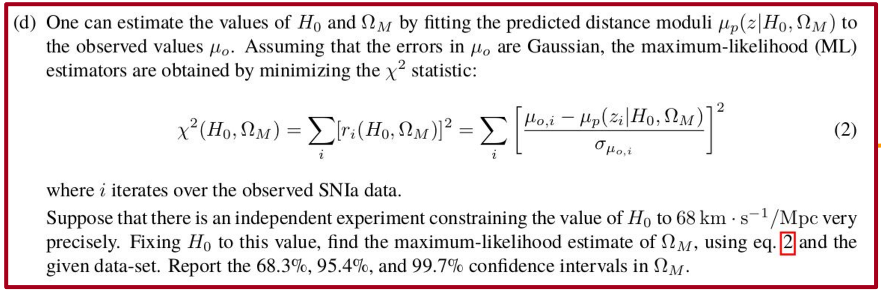


Compare against Paper



Stanford University

Replicate Nobel Prize Winning Discovery: SNIa Evidence for Expanding Universe (2/2)

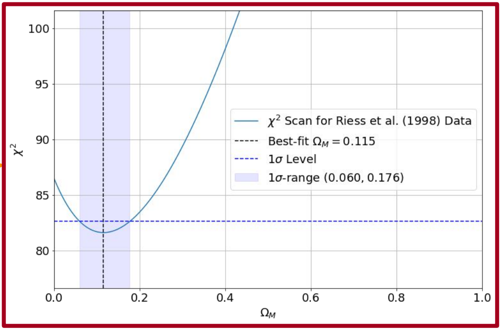


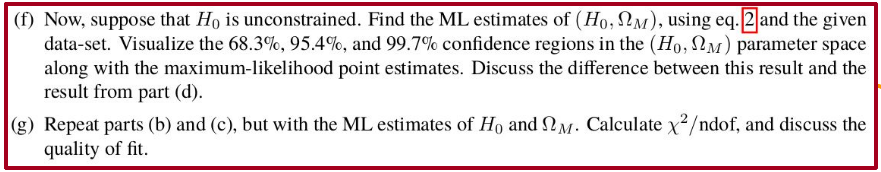
(d) One can estimate the values of H_0 and Ω_M by fitting the predicted distance moduli $\mu_p(z|H_0, \Omega_M)$ to the observed values μ_o. Assuming that the errors in μ_o are Gaussian, the maximum-likelihood (ML) estimators are obtained by minimizing the χ^2 statistic:

$$\chi^2(H_0, \Omega_M) = \sum_i [r_i(H_0, \Omega_M)]^2 = \sum_i \left[\frac{\mu_{o,i} - \mu_p(z_i|H_0, \Omega_M)}{\sigma_{\mu_{o,i}}}\right]^2 \tag{2}$$

where i iterates over the observed SNIa data.

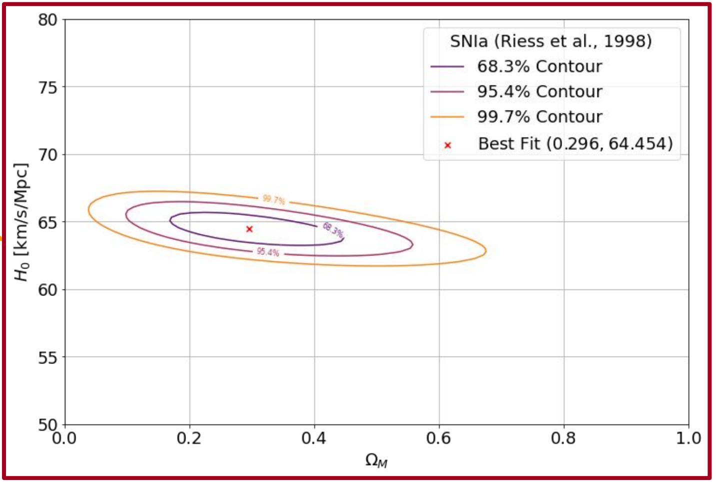
Suppose that there is an independent experiment constraining the value of H_0 to $68\,\mathrm{km} \cdot \mathrm{s}^{-1}/\mathrm{Mpc}$ very precisely. Fixing H_0 to this value, find the maximum-likelihood estimate of Ω_M, using eq. $\boxed{2}$ and the given data-set. Report the 68.3%, 95.4%, and 99.7% confidence intervals in Ω_M.



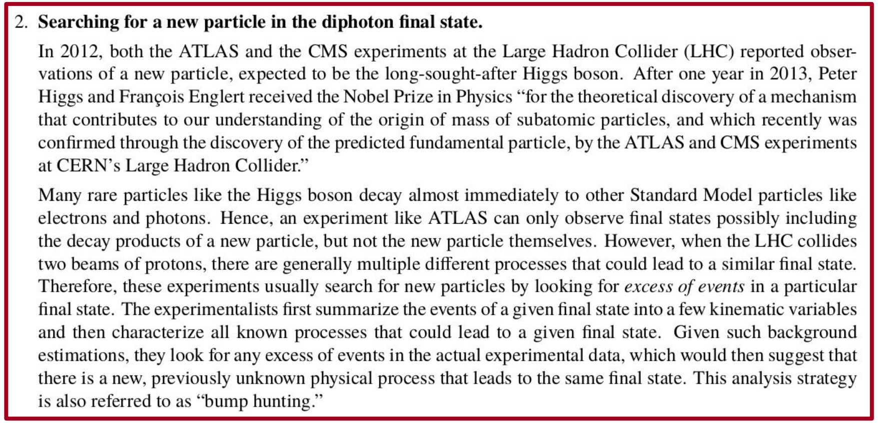


(f) Now, suppose that H_0 is unconstrained. Find the ML estimates of (H_0, Ω_M), using eq. $\boxed{2}$ and the given data-set. Visualize the 68.3%, 95.4%, and 99.7% confidence regions in the (H_0, Ω_M) parameter space along with the maximum-likelihood point estimates. Discuss the difference between this result and the result from part (d).

(g) Repeat parts (b) and (c), but with the ML estimates of H_0 and Ω_M. Calculate χ^2/ndof, and discuss the quality of fit.



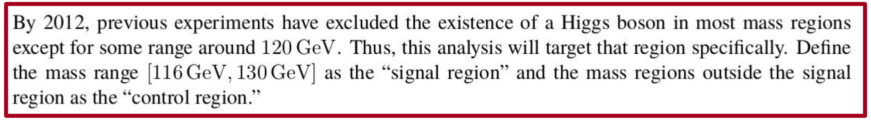
Replicate Nobel Prize Winning Discovery: Higgs Boson Discovery in Diphoton Channel



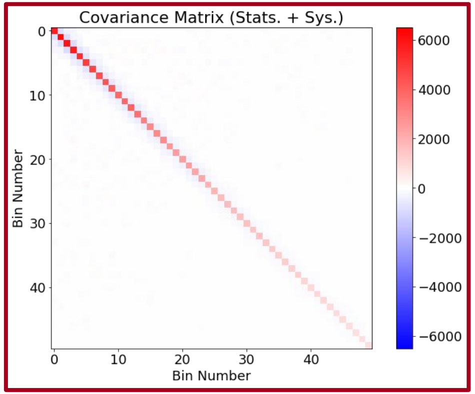
2. **Searching for a new particle in the diphoton final state.**

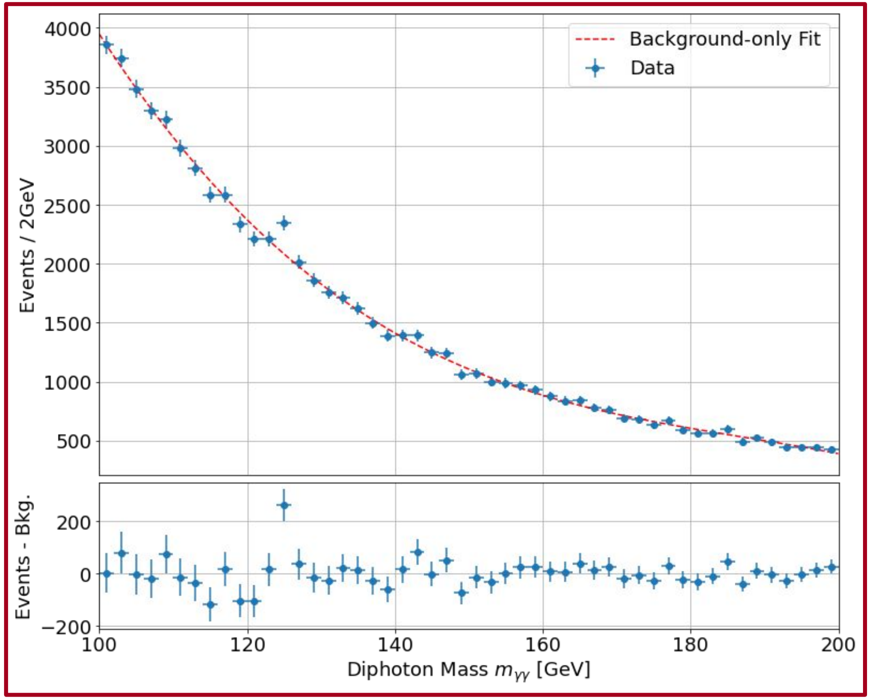
In 2012, both the ATLAS and the CMS experiments at the Large Hadron Collider (LHC) reported observations of a new particle, expected to be the long-sought-after Higgs boson. After one year in 2013, Peter Higgs and François Englert received the Nobel Prize in Physics "for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider."

Many rare particles like the Higgs boson decay almost immediately to other Standard Model particles like electrons and photons. Hence, an experiment like ATLAS can only observe final states possibly including the decay products of a new particle, but not the new particle themselves. However, when the LHC collides two beams of protons, there are generally multiple different processes that could lead to a similar final state. Therefore, these experiments usually search for new particles by looking for *excess of events* in a particular final state. The experimentalists first summarize the events of a given final state into a few kinematic variables and then characterize all known processes that could lead to a given final state. Given such background estimations, they look for any excess of events in the actual experimental data, which would then suggest that there is a new, previously unknown physical process that leads to the same final state. This analysis strategy is also referred to as "bump hunting."



By 2012, previous experiments have excluded the existence of a Higgs boson in most mass regions except for some range around 120 GeV. Thus, this analysis will target that region specifically. Define the mass range $[116\,\mathrm{GeV}, 130\,\mathrm{GeV}]$ as the "signal region" and the mass regions outside the signal region as the "control region."





Results & Lessons Learned

Student Reviews, Instructor Reflections

Overall Highlights from Student Reviews

Shamelessly cherry-picked 🙂

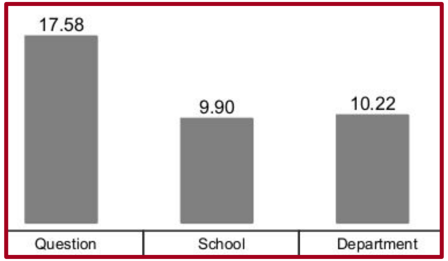
- “How to do actual science. To be more specific, **to be able to frame physics in a data-driven way and be able to extract meaningful results** and see when results are not meaningful through various tools in probability and statistics. … Also, being able to deal with real data in problem sets and through the final project have given me a lot of **applied skills in programming and data analysis that I am absolutely sure I'll use in the future**.”
- “The class covers topics that every experimental physicist should know, especially if you haven't taken any stat classes. **Also, if you're thinking about taking classes like CS 229 (Machine Learning), you should take this class before that.**”
- “Classes and concepts are presented intuitively **without losing mathematical formality.**”
- “The course delves into the philosophy which motivates real-world physics and managed to make me (a theorist) appreciate experimental frameworks in an entirely new light.”

More Specific Successes

Shamelessly cherry-picked 🙂

- “I thought this **(the basic probability) was absolutely necessary** and helped me make sense of probability theory, even after having taken the introductory course at Stanford.”
- “Supernovae Evidence for Acceleration of the Universe -- this problem required the **synthesis of a great number of skills** we acquired over the course and made me feel like a capable statistics beginner in physics.”
- “... my absolute favorite (problems) were the ones where we could **code things up and see how skills are applied in action.**”
- “I enjoyed this class because it taught stats from **a rigorous perspective**.”

Challenges & Remaining Tasks



Admittedly difficult issues 😐

- “How many hours per week on average did you spend on this course (including class meetings)?”
  - “One of the most time-consuming physics classes I've taken”
  - Nominally 3hrs / unit × 4 units = 12hrs
  - **This is a very ambitious, heavy course…**
- “for people who were **not as familiar with coding** as I was, some of the coding questions may have been a bit difficult”
  - Indeed! Major challenge if we were to deploy this material to a wider audience
- “The last few lectures were too fast.”
  - We often feel **rushed towards the end**, skipping/missing topics
- **Instructor biases towards particles physics & astrophysics examples**
  - Community feedback/input would be immensely helpful!

Thank You!

Please come talk to me during breaks, unconference, and so on!

Sanha Cheong (sanha@stanford.edu)
Ariel Schwartzman (sch@slac.stanford.edu)