# Statistics for your Analysis

USCMS Undergrad. Internship Program

August 12, 2022

Jieun Yoo

Today's talk will focus on understanding the uncertainties on scientific measurements and learning some statistical tools to help you interpret your data

## How did we discover the Higgs boson?

The Higgs boson can't be "discovered" by finding it somewhere but has to be created in a particle collision. Once created, it transforms – or "decays" – into other particles that can be detected in particle detectors.

Physicists look for traces of these particles in data collected by the detectors. The challenge is that these particles are also produced in many other processes, plus the Higgs boson only appears in about one in a billion LHC collisions. But careful statistical analysis of enormous amounts of data uncovered the particle's faint signal in 2012.

https://home.cern/science/physics/higgs-boson#:~:text=In%20the%20Higgs%20boson's%20case,existence%20of%20the%20Higgs%20field

# Outline

01 Introduction

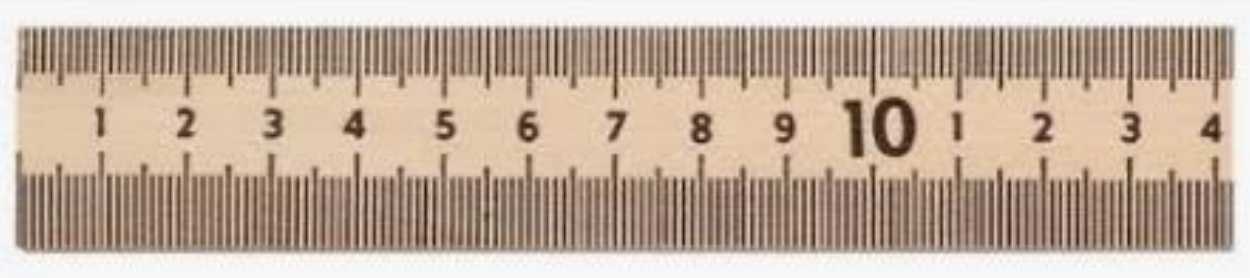
02 Measuring & Reporting Uncertainties

03 P-values, Sigma

04 Stat. & Systematic Uncertainties

05 Closing

# Measurement and Uncertainty




What's the smallest division on a meter stick?

What kinds of measurements make sense?

# Instrument Uncertainty

**STANDARD FORM FOR STATING UNCERTAINTIES**

The standard form for reporting a measurement of a physical quantity $x$ is

$$(\text{measured value of } x) = x_{\text{best}} \pm \delta x,$$

where

$$x_{\text{best}} = (\text{best estimate for } x)$$

and

$$\delta x = (\text{uncertainty or error in the measurement}). \qquad [\text{See (2.3)}]$$

This statement expresses our confidence that the correct value of $x$ probably lies in (or close to) the range from $x_{\text{best}} - \delta x$ to $x_{\text{best}} + \delta x$.

Taylor (1997)

Rule of thumb: uncertainty when using a measuring device with a scale = smallest increment/2
So, for a meter stick the uncertainty would be 1 mm/2 = 0.5 mm or 0.05 cm

Something measured as 5.5 cm could actually be anywhere 5.45 cm <= x <= 5.55 cm

# Standard Error (of the Mean)

- Make a measurement N number of times
- "Mean" is the same as Average
- S.D. is a measure of how far each measurement is from the mean.

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i)^2}$$

**Table 4.2.** Calculation of the standard deviation. Example from Taylor (1997)

| Trial number $i$ | Measured value $x_i$ | Deviation $d_i = x_i - \bar{x}$ | Deviation squared $d_i^2$ |
|---|---|---|---|
| 1 | 71 | −0.8 | 0.64 |
| 2 | 72 | 0.2 | 0.04 |
| 3 | 72 | 0.2 | 0.04 |
| 4 | 73 | 1.2 | 1.44 |
| 5 | 71 | −0.8 | 0.64 |
| | $\Sigma x_i = 359$ | $\Sigma d_i = 0.0$ | $\Sigma d_i^2 = 2.80$ |

$$\bar{x} = 359/5 = 71.8$$

$$\sigma_x^2 = \frac{1}{N}\Sigma d_i^2 = \frac{2.80}{5} = 0.56$$

$$\sigma_x \approx 0.7$$

Standard Error of the Mean (Standard Deviation of the Mean)

$$\delta x = \sigma_{\bar{x}} = \sigma_x/\sqrt{N}$$

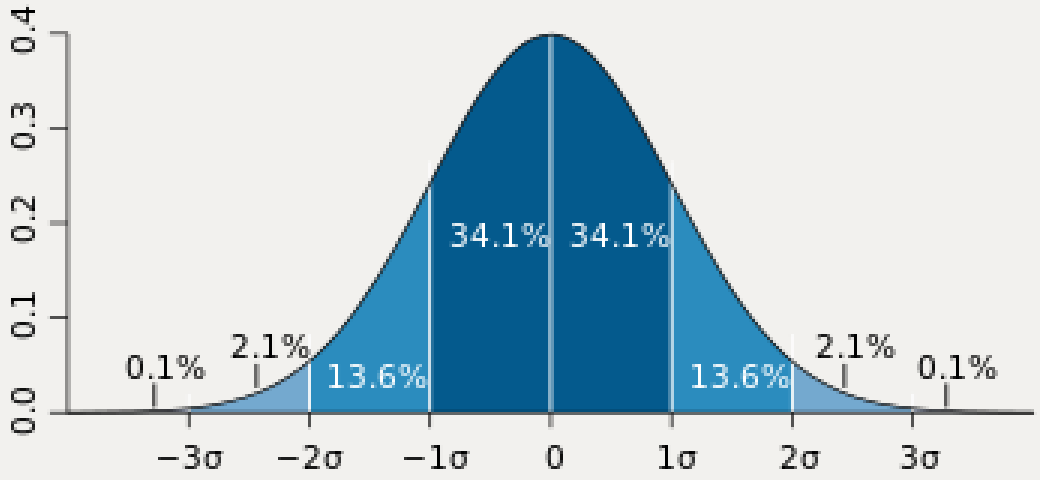
# Square Root Rule for Counting Experiments

Example:

- a radioactive material decays at an average rate
- We count number of decays in some time T
- We can make a definitive count
- The uncertainty is how well this approaches the true avg. number

(average number of events in time $T$) $= \nu \pm \sqrt{\nu}$.

# Normal Distribution (Gaussian distribution)




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

https://en.wikipedia.org/wiki/Normal_distribution

# Hypotheses, p-values and more!

- $H_0$ : Null Hypothesis – Background only
- $H_1$: Alternative Hypothesis – Background + Signal

- Choose a test statistic
  - "a single number that quantifies the entire experiment"
  - Ex: # of events, chi-square, ratio of log likelihoods, BDT score

- Choose a significance level alpha
- Alpha is defined ahead of the time; it doesn't come from data
- If p-value is <= alpha, then you can reject $H_0$

- Get p-value from data

# P value

- Definition: "the probability of obtaining test results at least as extreme as the result actually observed"

- A smaller p-value means more evidence in favor or rejecting the null hypothesis

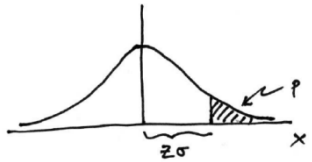
- You can easily convert p-values to sigma https://rdrr.io/snippets/
  - In R: pnorm(5) -> 0.9999997
  - (1-pnorm(5))-> 2.866516e-07
- Or, in ROOT

```
root [2] sigma=ROOT::Math::normal_quantile_c(0.0000002866516,1)
(double) 5.0000000
```

*From Cowan's lectures*:
https://www.pp.rhul.ac.uk/~cowan/stat/stat_7.pdf

Significance from $p$-value

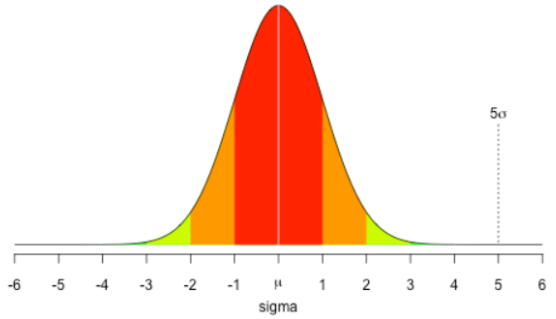
Often define significance $Z$ as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same $p$-value.




$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$ `1 - TMath::Freq`

$$Z = \Phi^{-1}(1-p)$$ `TMath::NormQuantile`

G. Cowan Lectures on Statistical Data Analysis




https://sasandr.wordpress.com/2012/07/14/god-particle-5-sigma-and-p-value/

# Look elsewhere effect

We looked at the “local” p-value.

The chance of a 5 sigma fluctuation in one bin is small, but histograms have many bins!

What if we looked elsewhere?

What if our result was just due to random fluctuations?

https://cms.cern/physics/cms-higgs-search/look-elsewhere-effect

# Blind Analysis

- Clever Hans Effect
- Blind Analysis – prevent experimenters from knowing the result until the analysis is complete
- Example technique: Hidden Signal Box
  - Hide a subset of the data with potential signal until the analysis is complete

https://www.britannica.com/topic/Clever-Hans

https://www.annualreviews.org/doi/pdf/10.1146/annurev.nucl.55.090704.151521

# Evidence or Discovery?

3.5 Sigma – for “Evidence”

5 Sigma – for “Discovery”

# Caveats

" In 2011 the OPERA collaboration produced a measurement of neutrino travel times from CERN to Gran Sasso which appeared smaller by 6σ than the travel time of light in vacuum[15]. The effect spurred lively debates, media coverage, checks by the ICARUS experiment and dedicated beam runs. It was finally understood to be due to a large source of systematic uncertainty – a loose cable[16] "

T. Dorigo, Extraordinary Claims, 2014

https://indico.cern.ch/event/277650/contributions/629796/attachments/505859/698408/Extraordinary_Claims.pdf

# Understanding a reported measurement

Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC

Results are presented from searches for the standard model Higgs boson in proton-proton collisions at $\sqrt{s} = 7$ and 8 TeV in the Compact Muon Solenoid experiment at the LHC, using data samples corresponding to integrated luminosities of up to 5.1 $fb^{-1}$ at 7 TeV and 5.3 $fb^{-1}$ at 8 TeV. The search is performed in five decay modes: $\gamma\gamma$, ZZ, $W^+W^-$, $\tau^+\tau^-$, and $b\bar{b}$. An excess of events is observed above the expected background, with a local significance of 5.0 standard deviations, at a mass near 125 GeV, signalling the production of a new particle. The expected significance for a standard model Higgs boson of that mass is 5.8 standard deviations. The excess is most significant in the two decay modes with the best mass resolution, $\gamma\gamma$ and ZZ; a fit to these signals gives a mass of $125.3 \pm 0.4\,(\text{stat.}) \pm 0.5\,(\text{syst.})$ GeV. The decay to two photons indicates that the new particle is a boson with spin different from one.

$$125.3 \pm 0.4\,(\text{stat.}) \pm 0.5\,(\text{syst.})\ \text{GeV}$$

https://arxiv.org/pdf/1207.7235.pdf

# Statistical Uncertainties

- They arise due to stochastic fluctuations since we make a limited number of observations

- Think back to our case of a meterstick: we can only make a finite number of measurements

- How does this apply in physics analyses? We have limited statistics

- With the HL-LHC upgrade, we will have higher luminosity -> higher stats!

# Systematic Uncertainties

Cause: uncertainties related to our detectors, our assumptions, our theoretical models

Our meterstick example:

- the ruler could be deformed
- The length of the ruler could change with temperature

Our physics analysis:

- Imperfect calibration of measuring devices
- Object energy resolution
- Reconstruction efficiencies
- Assumptions made in our Monte Carlo models

# Summary

UNDERSTANDING A MEASUREMENT REQUIRES UNDERSTANDING UNCERTAINTIES

STATISTICAL ANALYSIS IS NEEDED TO INTERPRET YOUR RESULTS

MANY SOPHISTICATED TECHNIQUES ARE NEEDED TO ANALYZE A COMPLEX DETECTOR LIKE CMS

# References


John R . Taylor, An introduction to error analysis, 2nd edition, University Science Books, 1997

https://www.physi.uni-heidelberg.de/~nberger/teaching/ws12/statistics/Lecture11.pdf

https://indico.cern.ch/event/508168/contributions/2028747/attachments/1307803/1962991/Statistical-Reasoning-HASCO16.pdf

https://www.slac.stanford.edu/econf/C030908/papers/TUAT004.pdf

https://indico.cern.ch/event/277650/contributions/629796/attachments/505859/698408/Extraordinary_Claims.pdf

# Backup

# Standard Deviation – alternative definition

Taylor (1997)

Unfortunately, the standard deviation has an alternative definition. There are theoretical arguments for replacing the factor $N$ in (4.6) by $(N - 1)$ and defining the standard deviation $\sigma_x$ of $x_1, \ldots, x_N$ as

$$\sigma_x = \sqrt{\frac{1}{N-1}\sum d_i^2} = \sqrt{\frac{1}{N-1}\sum (x_i - \bar{x})^2}. \qquad (4.9)$$

I will not try here to prove that definition (4.9) of $\sigma_x$ is better than (4.6), except to say that the new "improved" definition is obviously a little larger than the old one (4.6) and that (4.9) corrects a tendency for (4.6) to understate the uncertainty in the measurements $x_1, \ldots, x_N$, especially if the number of measurements $N$ is small. This tendency can be understood by considering the extreme (and absurd) case that $N = 1$ (that is, we make only one measurement). Here, the average $\bar{x}$ is equal to our one reading $x_1$, and the one deviation is automatically zero. Therefore, the definition (4.6) gives the absurd result $\sigma_x = 0$. On the other hand, the definition (4.9) gives 0/0; that is, with definition (4.9), $\sigma_x$ is undefined, which correctly reflects our total ignorance of the uncertainty after just one measurement. The definition (4.6) is sometimes called the *population standard deviation* and (4.9) the *sample standard deviation.*

The difference between the two definitions (4.6) and (4.9) is almost always numerically insignificant. You should always repeat a measurement many times (at

# Error Propagation

If various quantities $x, \ldots, w$ are measured with small uncertainties $\delta x, \ldots, \delta w$, and the measured values are used to calculate some quantity $q$, then the uncertainties in $x, \ldots, w$ cause an uncertainty in $q$ as follows:
If $q$ is the sum and difference, $q = x + \cdots + z - (u + \cdots + w)$, then

$$\delta q \begin{cases} = \sqrt{(\delta x)^2 + \cdots + (\delta z)^2 + (\delta u)^2 + \cdots + (\delta w)^2} \\ \quad \text{for independent random errors;} \\ \leq \delta x + \cdots + \delta z + \delta u + \cdots + \delta w \\ \quad \text{always.} \end{cases} \qquad \text{(p. 60)}$$

If $q$ is the product and quotient, $q = \dfrac{x \times \cdots \times z}{u \times \cdots \times w}$, then

$$\frac{\delta q}{|q|} \begin{cases} = \sqrt{\left(\dfrac{\delta x}{x}\right)^2 + \cdots + \left(\dfrac{\delta z}{z}\right)^2 + \left(\dfrac{\delta u}{u}\right)^2 + \cdots + \left(\dfrac{\delta w}{w}\right)^2} \\ \quad \text{for independent random errors;} \\ \leq \dfrac{\delta x}{|x|} + \cdots + \dfrac{\delta z}{|z|} + \dfrac{\delta u}{|u|} + \cdots + \dfrac{\delta w}{|w|} \\ \quad \text{always.} \end{cases} \qquad \text{(p. 61)}$$

If $q = Bx$, where $B$ is known exactly, then

$$\delta q = |B| \, \delta x. \qquad \text{(p. 54)}$$

If $q$ is a function of one variable, $q(x)$, then

$$\delta q = \left| \frac{dq}{dx} \right| \delta x. \qquad \text{(p. 65)}$$

If $q$ is a power, $q = x^n$, then

$$\frac{\delta q}{|q|} = |n| \frac{\delta x}{|x|}. \qquad \text{(p. 66)}$$

Taylor (1997)

# Chi-Squared

$$\chi^2 \equiv \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} + \cdots + \frac{(x_\nu - \mu_\nu)^2}{\sigma_\nu^2} = \sum_{i=1}^{\nu} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$