Statistics

Jonas Rademacker at TESHEP 2024

Statistics,Probability and Physics

$$i\hbar\frac{\partial}{\partial t}\psi = -\frac{\hbar^2}{2m}\nabla^2\psi + V\psi$$

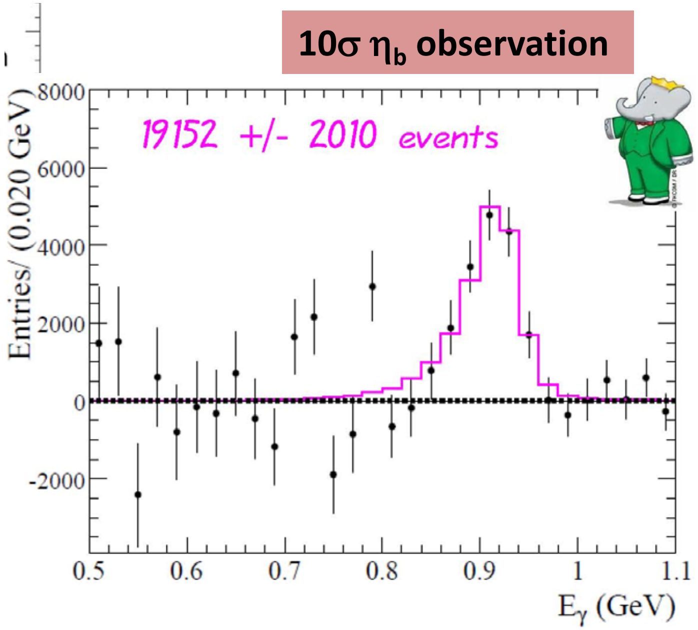
Quantum Mechanics

(a different, fundamental meaning to probability)

Thermodynamics

Probability, law of large numbers, combinatorics

10σ η_b observation



Interpretation of data

measurement errors, statistical fluctuations, Central Limit Theorem, confirming & rejecting theories, what constitutes a discovery?

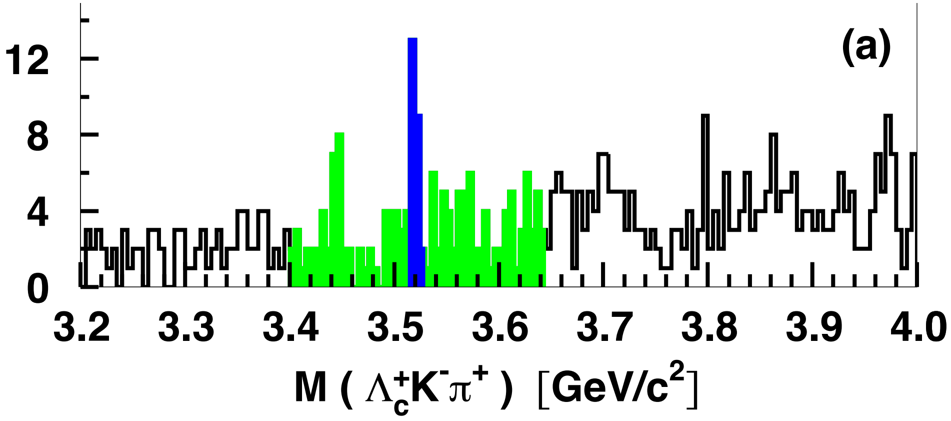
For a physics Masters/Ph.D….

- **You'll be looking at and interpreting a lot of data.**
- **You'll deal with a few basic distributions**
  - **Gaussian, Poisson, binomial, ... (and possibly a few others that you'll pick up as you go along)**
  - **You'll deal with error estimates and error matrices**
- **You'll measure parameters doing likelihood and χ^2 fits**
  - **You'll need to translate physics into PDF's**
  - **You'll interpret the fit result: what's the error? Is it a discovery? Are the data consistent with the PDF?**

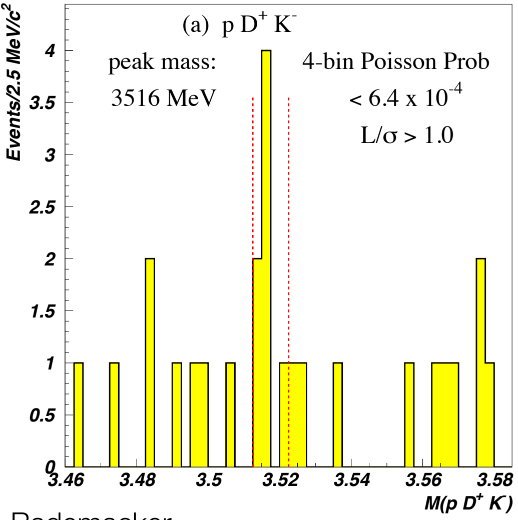
A Ξ_{cc} at 3.5 GeV?

SELEX see it twice

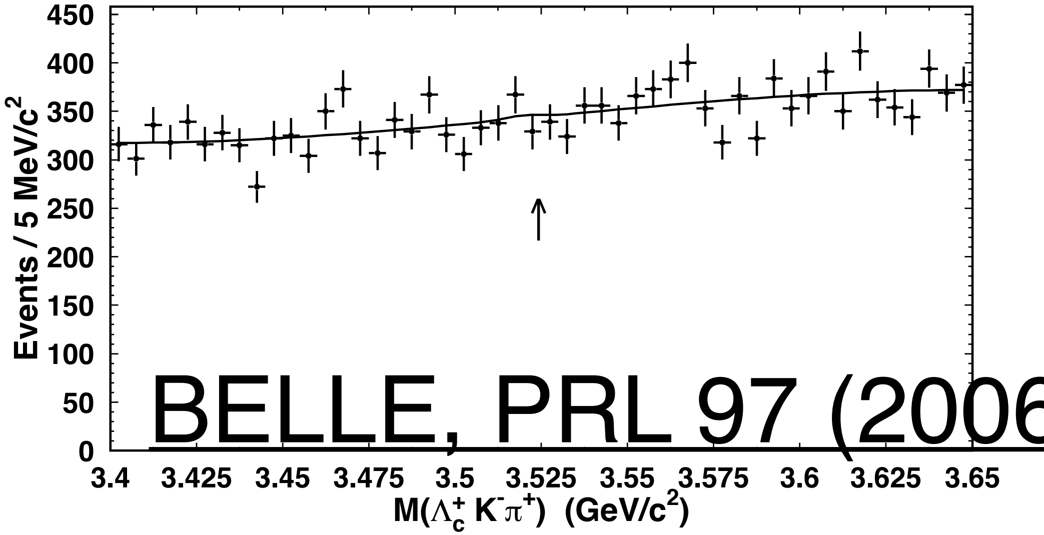
SELEX 2002



SELEX 2005



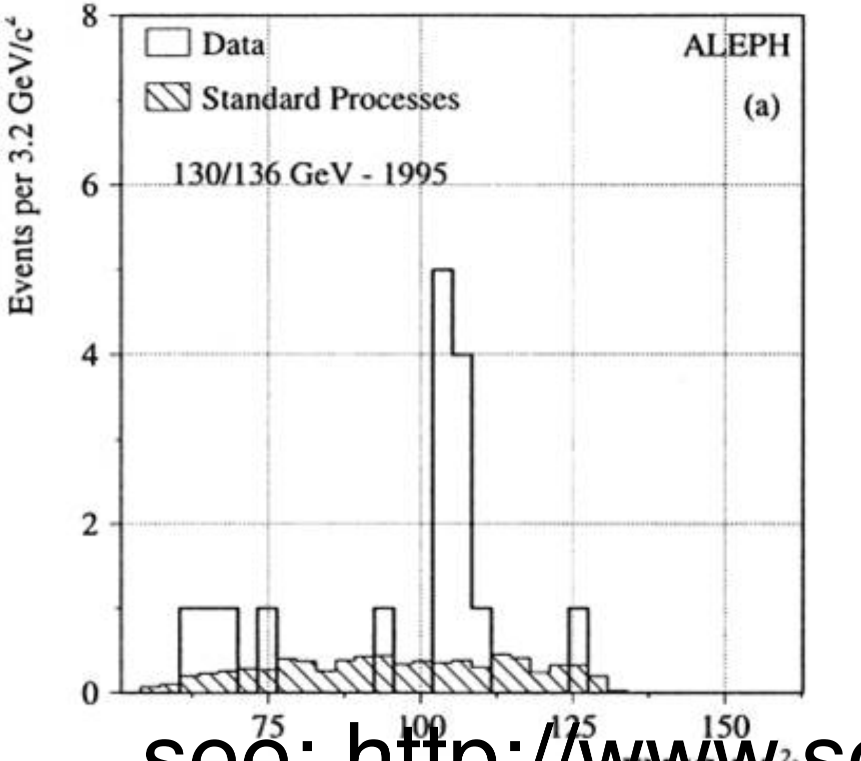
FOCUS, BaBar, BELLE, LHCb don't



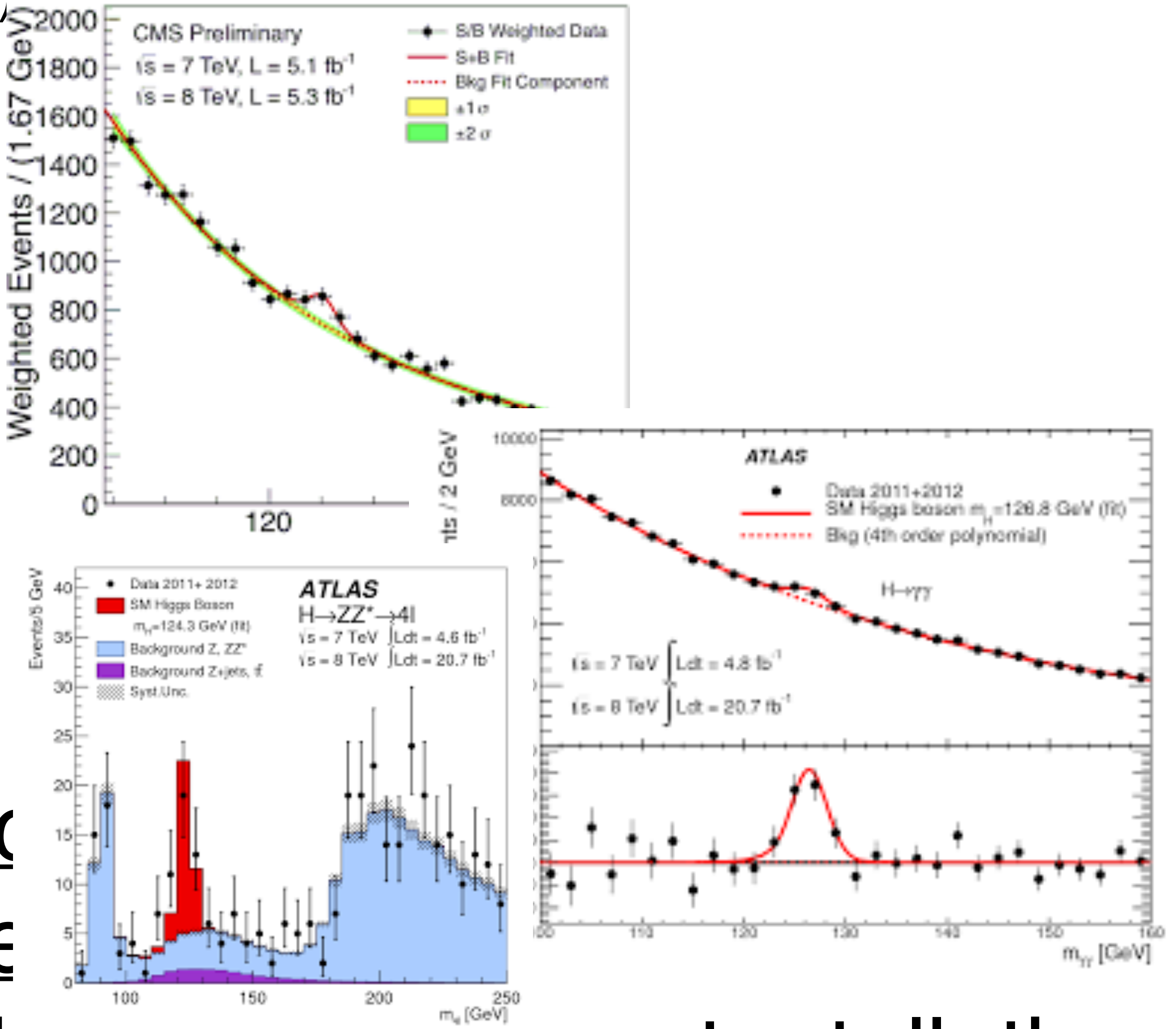
BELLE, PRL 97 (2006

Higgs: true or false?

false Higgs (ALEPH/LEP 1996)



real Higgs (LHC 1996)

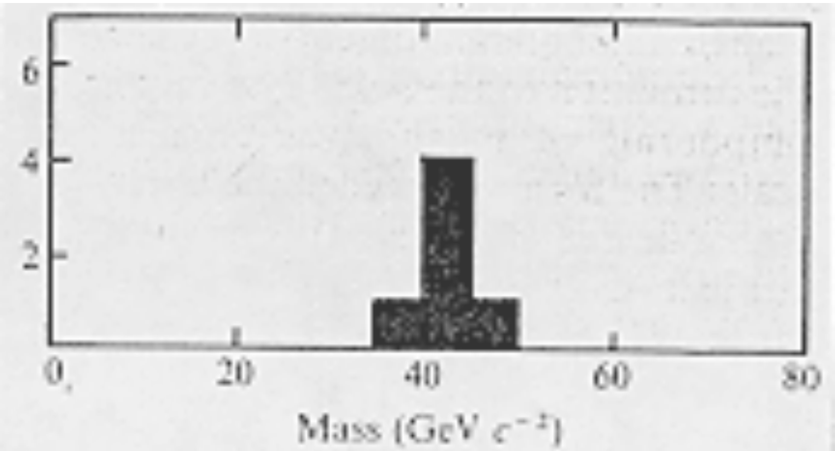


see: http://www.sc
a_quantum_diarie
true_and_false_discoveries_how_to_tell_them

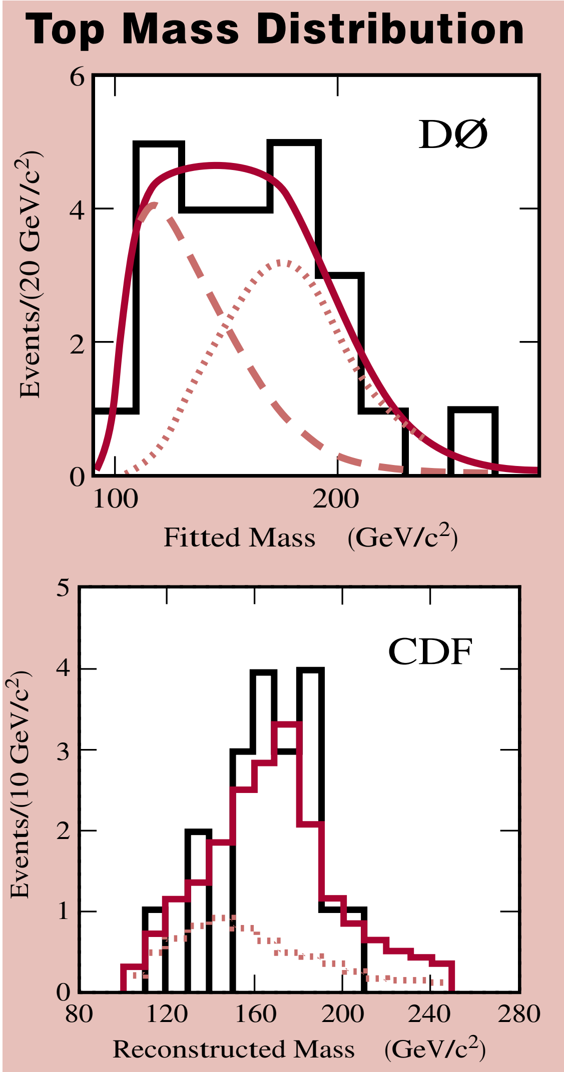
True and False

True top (1996)

False top (1985)

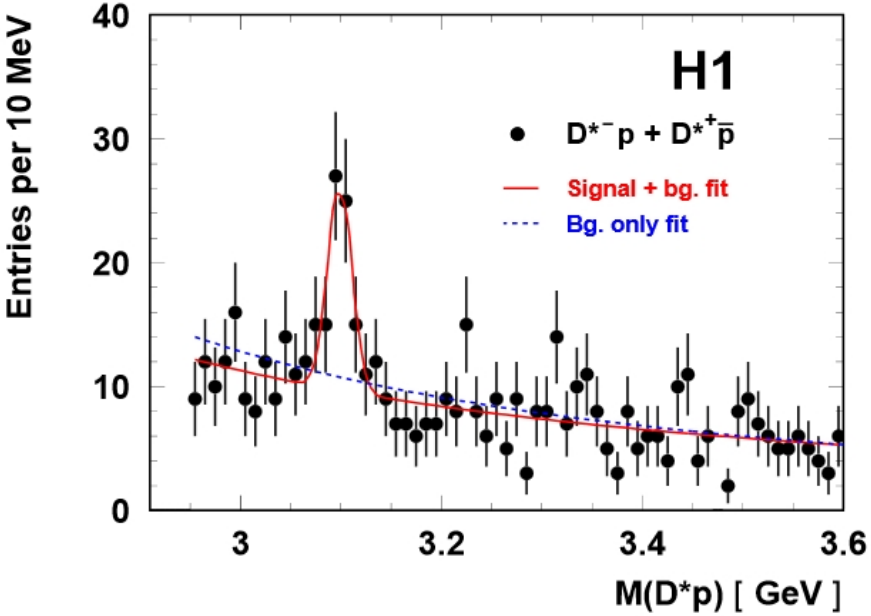


Distribution of measured top quark mass.

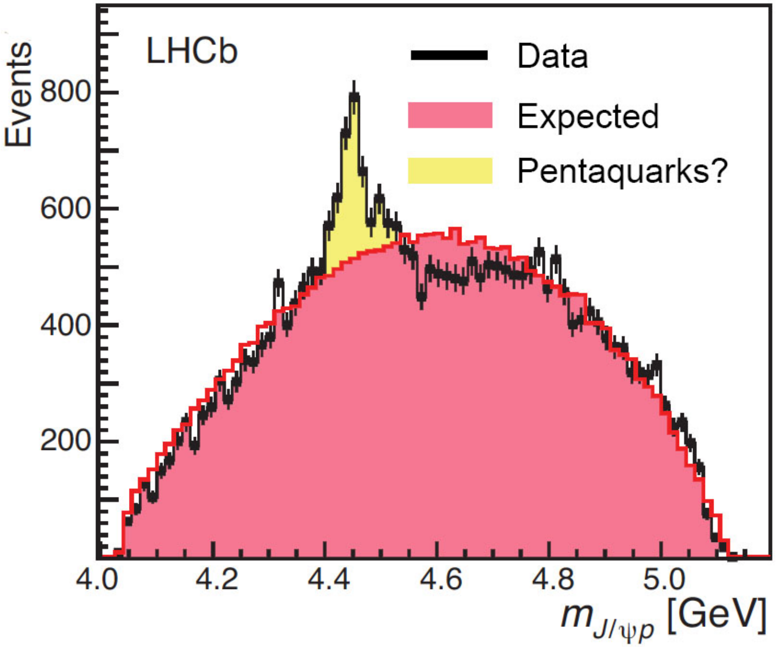

Top Mass Distribution

True & False: Pentaquark

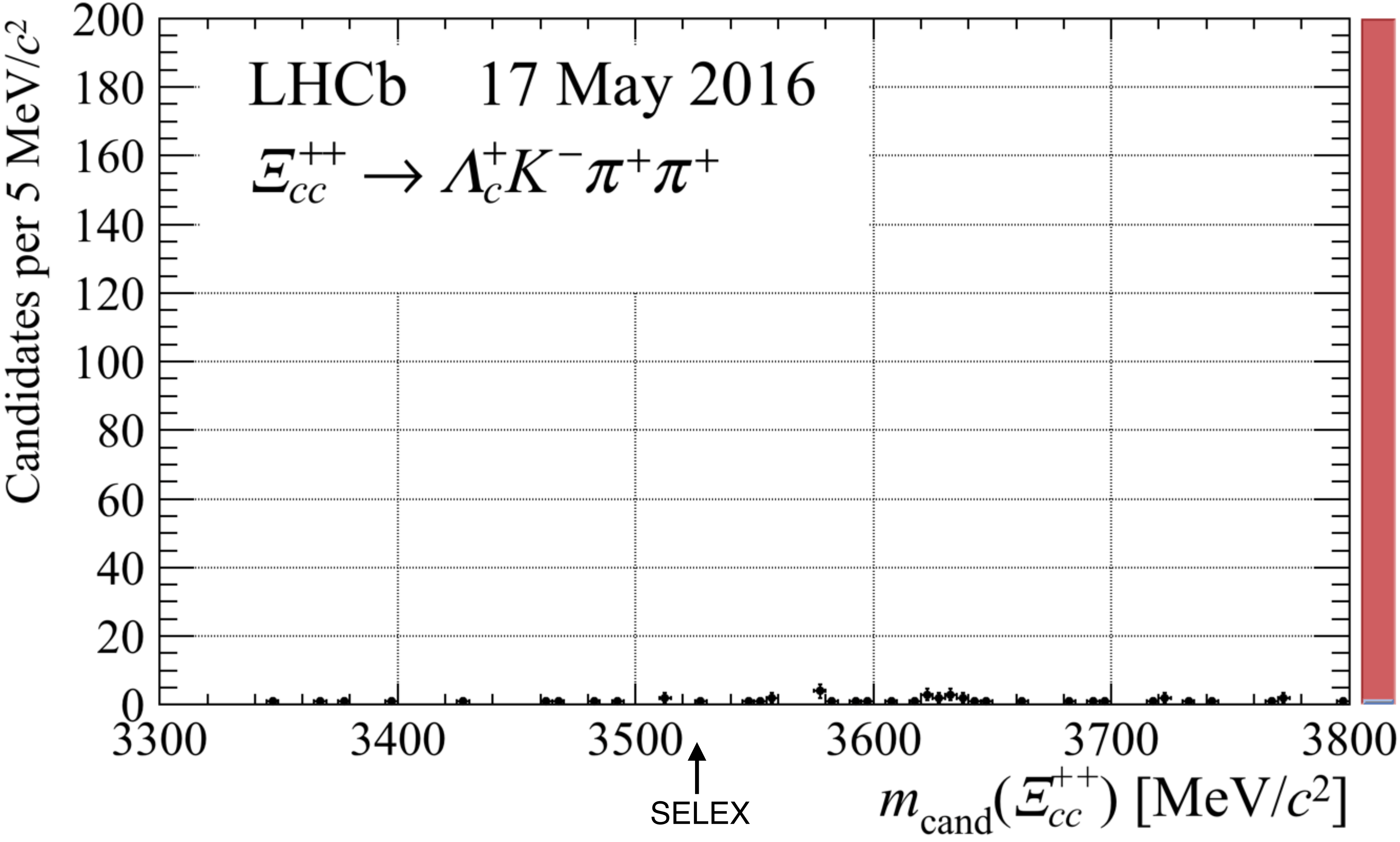
false, 2004, H1 (DESY)



true (LHCb, 2015)



When did this become a discovery?



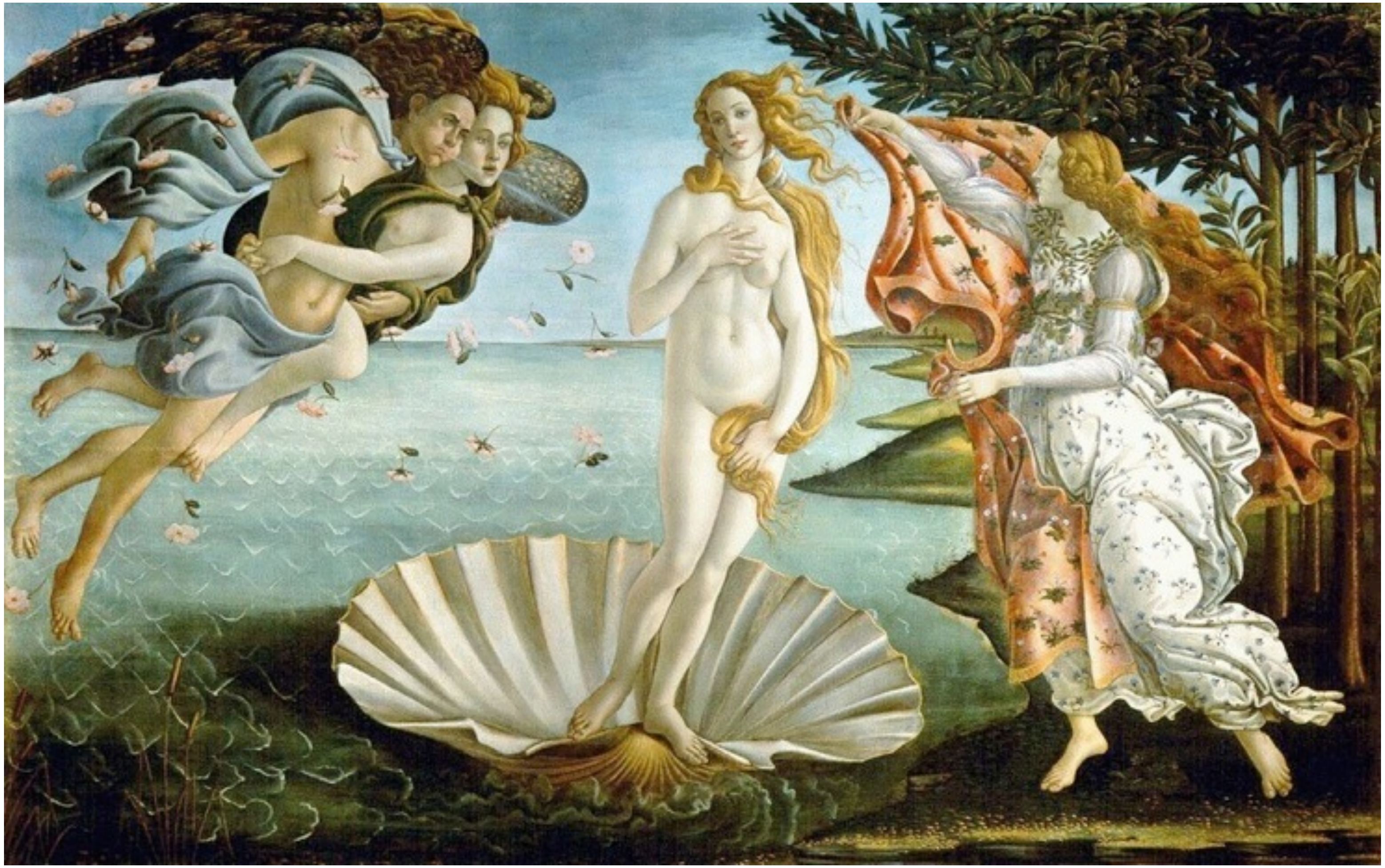
Discoveries...

- Particle physics is rife with false hints of discoveries - even the Higgs was seen and unseen at several energies before the LHC had its famous 5σ discovery.

- The problem: Nature does not allow us a direct view on its fundamental parameters.

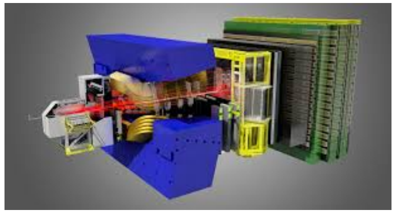
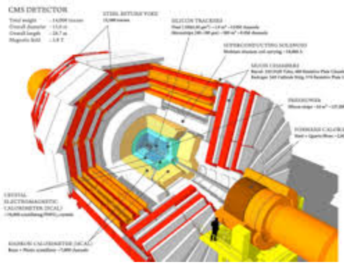
What we want

$\mathscr{L} =$



The Birth of Venus by Sandro Botticelli, c. 1482–1486. tempera on canvas, 172.5 × 278.5 cm, Uffizi, Florence

What we get



CMS DETECTOR
STEEL RETURN YOKE
SILICON TRACKERS
SUPERCONDUCTING SOLENOID
MUON CHAMBERS
PRESHOWER
FORWARD CALORIMETER
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
HADRON CALORIMETER (HCAL)

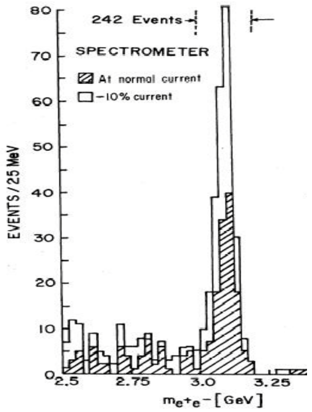
Statistics and Measurements

- Each measurement is messed up by millions of little perturbations that we cannot possibly all take into account, or even know about, individually.

- Statistics is the tool that allows us to separate the effect of those fluctuations from the underlying data. And it provides us with tools that tell us how confident we should be in our measurements.

- After this lecture, you won't discover a false Ξ_{cc} (OK, it's too late for that anyway) or a false Z'. I hope. Discover something surprising, and real!

Roadmap



What do I see?

Describing Data

What do I expect?

Probability and probability distributions, Probability density functions

Central Limit Theorem

Is what I see compatible with what I expect?

Discoveries
Confidence Levels
Hypothesis testing
Fitting

Monte Carlo simulation

Books

- R. J. Barlow: “Statistics”, John Wiley & Sons, ISBN 0-471-92295-1.

- Louis Lyons: “Statistics for nuclear and particle physicists”, Cambridge University Press, ISBN 0–521–37934–2

- Frederick James: “Statistical Methods in Experimental Physics”, World Scientific, ISBN 981-270-527-9 (pbk).

Problems

Problem sheets:

https://tinyurl.com/TeshepProblems

Code (Jupyter Notebooks):

https://tinyurl.com/TeshepStatCode

Problems

Problem sheets:

https://tinyurl.com/TeshepProblems

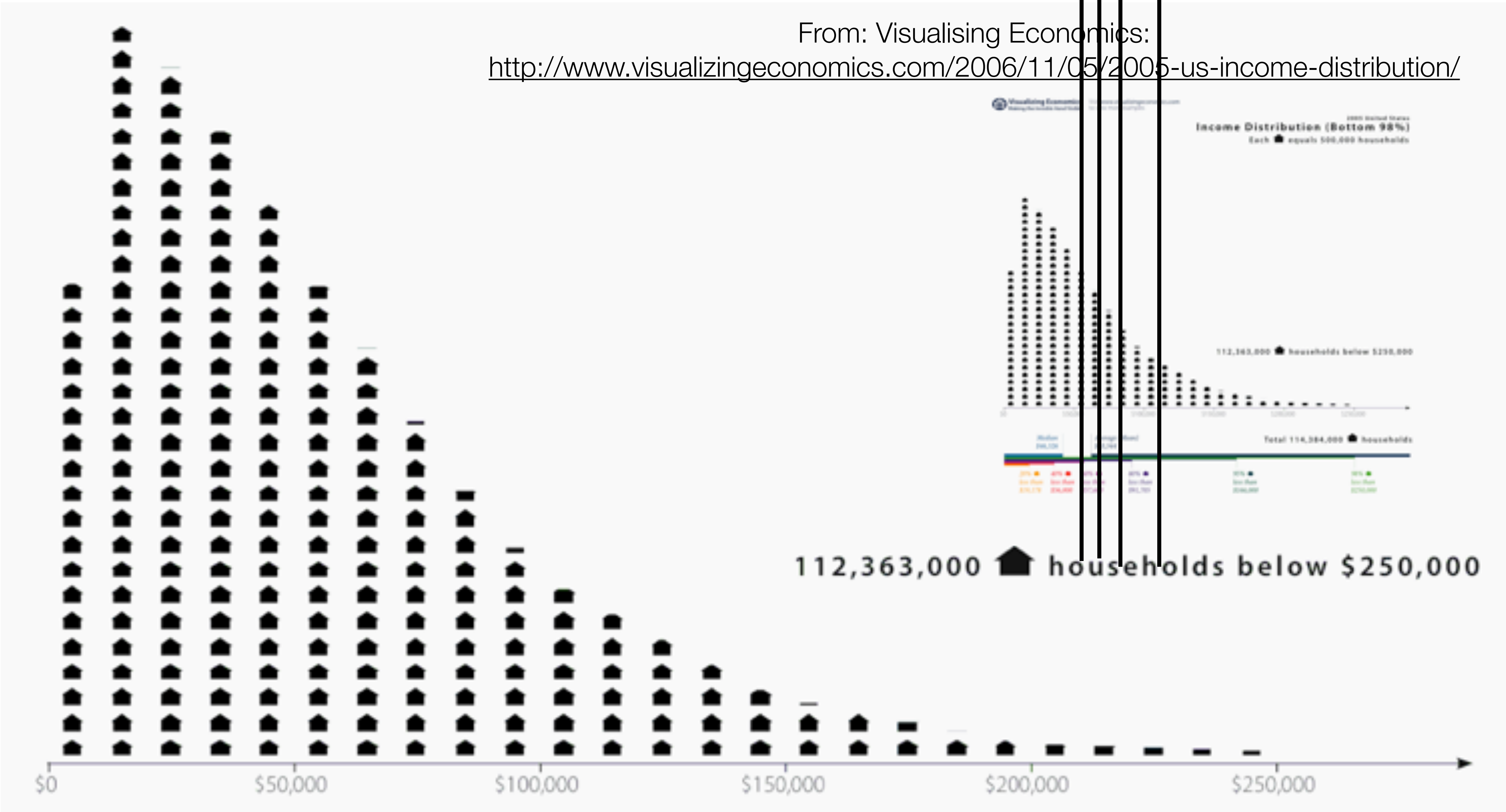
Code (Jupyter Notebooks):

https://tinyurl.com/TeshepStatCode

Describing data with numbers

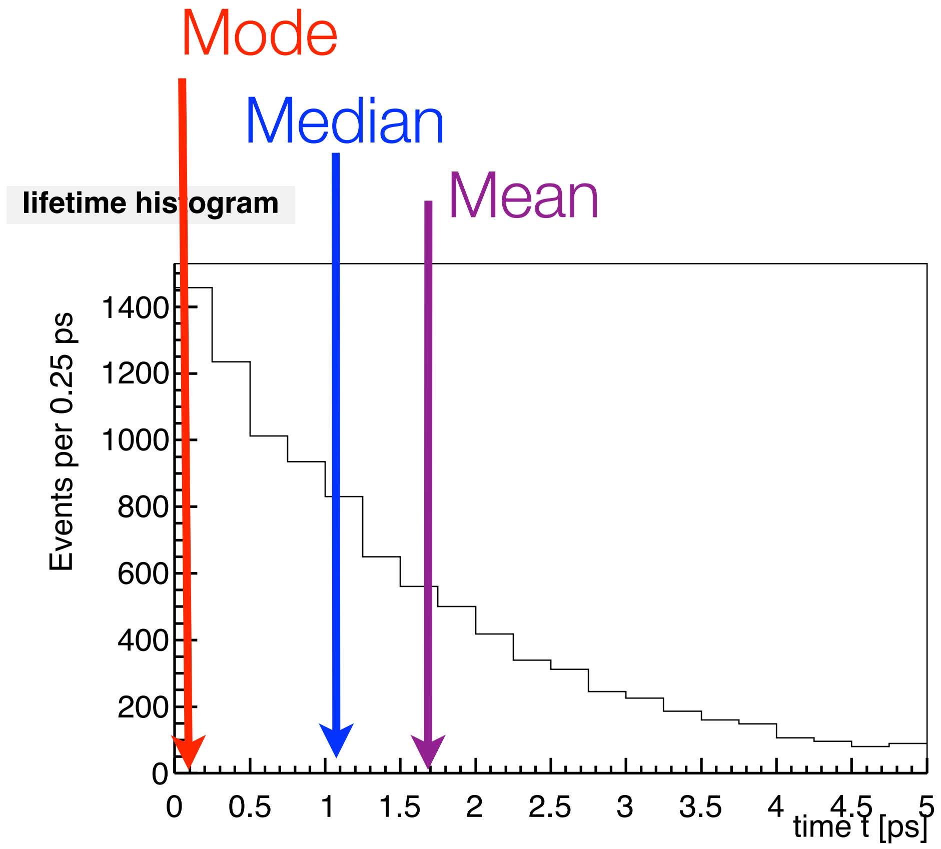
- **How do we describe a set of measurements with just a couple of characteristic, meaningful numbers?**

Annual Income



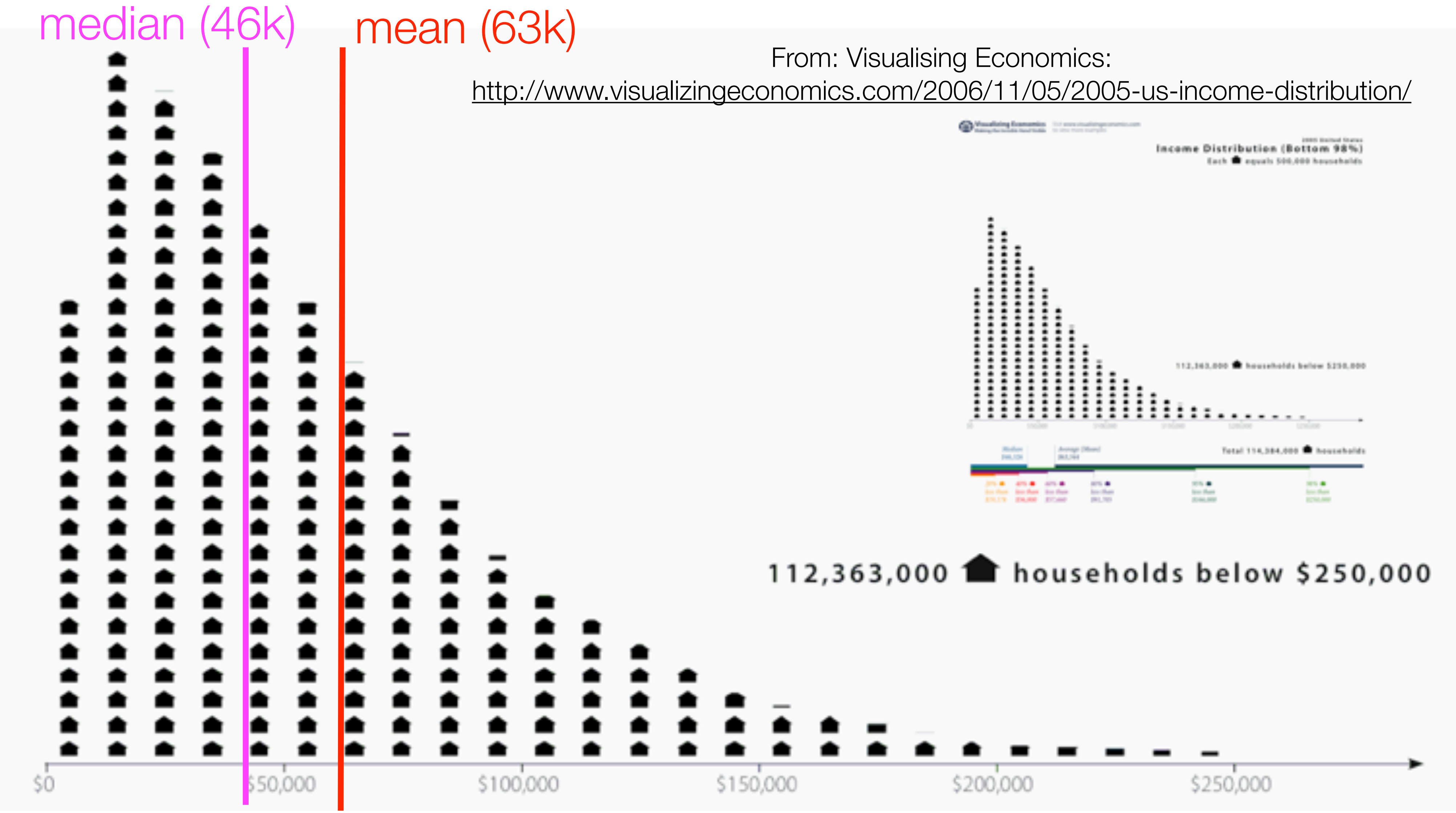
From: Visualising Economics:
http://www.visualizingeconomics.com/2006/11/05/2005-us-income-distribution/

Central Values



- **Mode: highest population**
- **Median: As many events below as above.**
- **Arithmetic Mean: $(1/N)\ \Sigma_{i=1,N}\ x_i$**

Annual Income

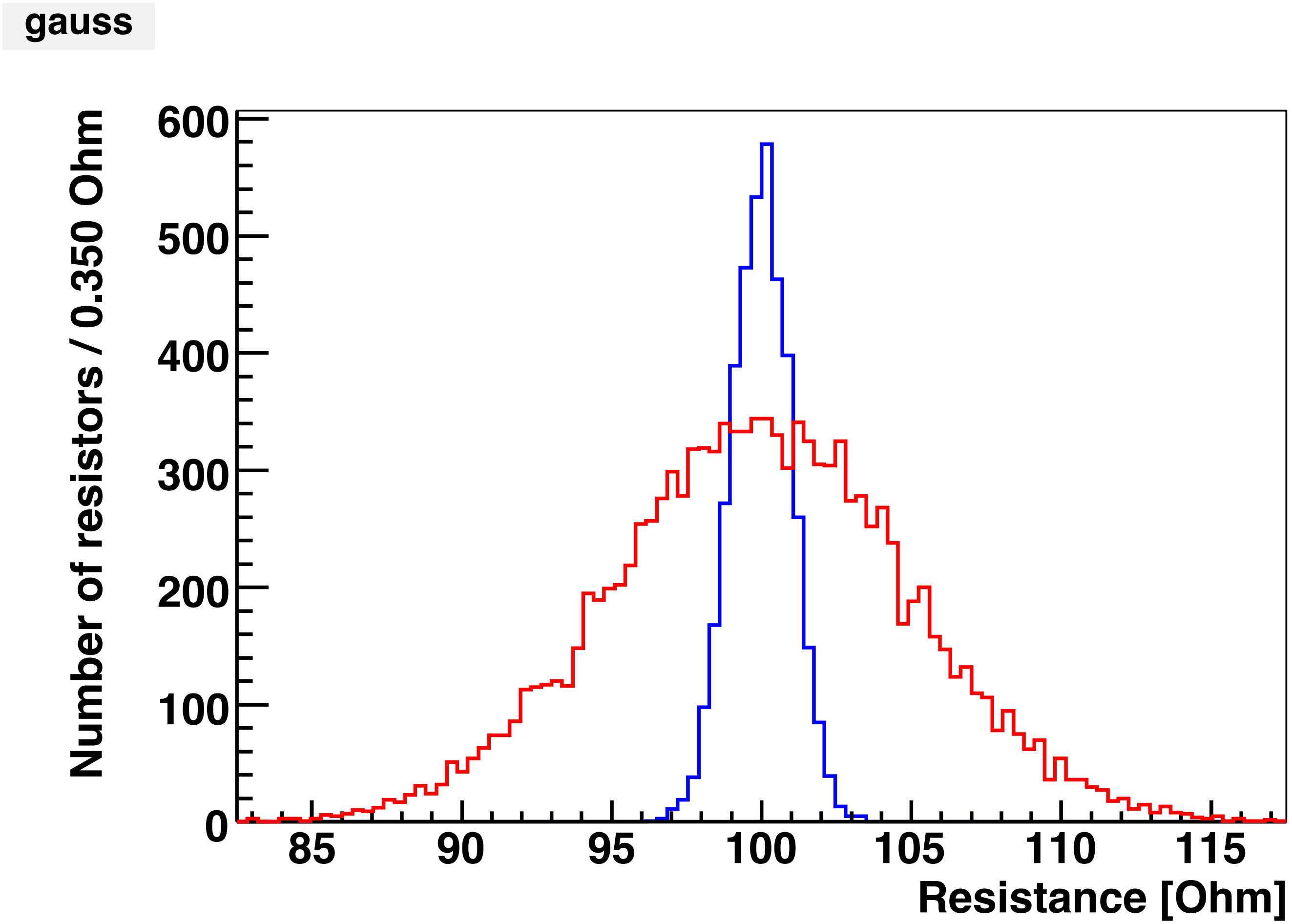


From: Visualising Economics: http://www.visualizingeconomics.com/2006/11/05/2005-us-income-distribution/

Mean

- For all practical purposes we will usually use the arithmetic mean: $(1/N)\ \Sigma_{i=1,N}\ x_i$
- Motivated to a large degree by its friendly mathematical properties.
- But other central values, other means exist (see also harmonic, geometric, etc) and they have their uses.

Width


gauss
Number of resistors / 0.350 Ohm
600
500
400
300
200
100
0
85
90
95
100
105
110
115
Resistance [Ohm]

Variance

- **We could calculate the total difference from the mean:**

  $d = \Sigma_{i=1,N} (x_i - \bar{x})$ but that's zero by the definition of the mean (check!)

- **The variance is the *average* (difference)^2 from the mean, the variance:**

- **$V \equiv \overline{(x - \bar{x})^2} = 1/N \, \Sigma_{i=1,N} (x_i - \bar{x})^2$**

Calculating the Variance

$$V = \overline{x^2} - \overline{x}^2$$

Home work:
verify this

- In words: The variance is equal to

  THE MEAN OF THE SQUARES

  MINUS

  THE SQUARE OF THE MEAN

- You'll always get the order of the terms right if you imagine a wide distribution centered at zero. $\overline{x}^2$ would zero, $\overline{x^2}$ positive and large, and the overall variance must not be negative.

Standard Deviation

- The Standard Deviation is the square-root of the variance:

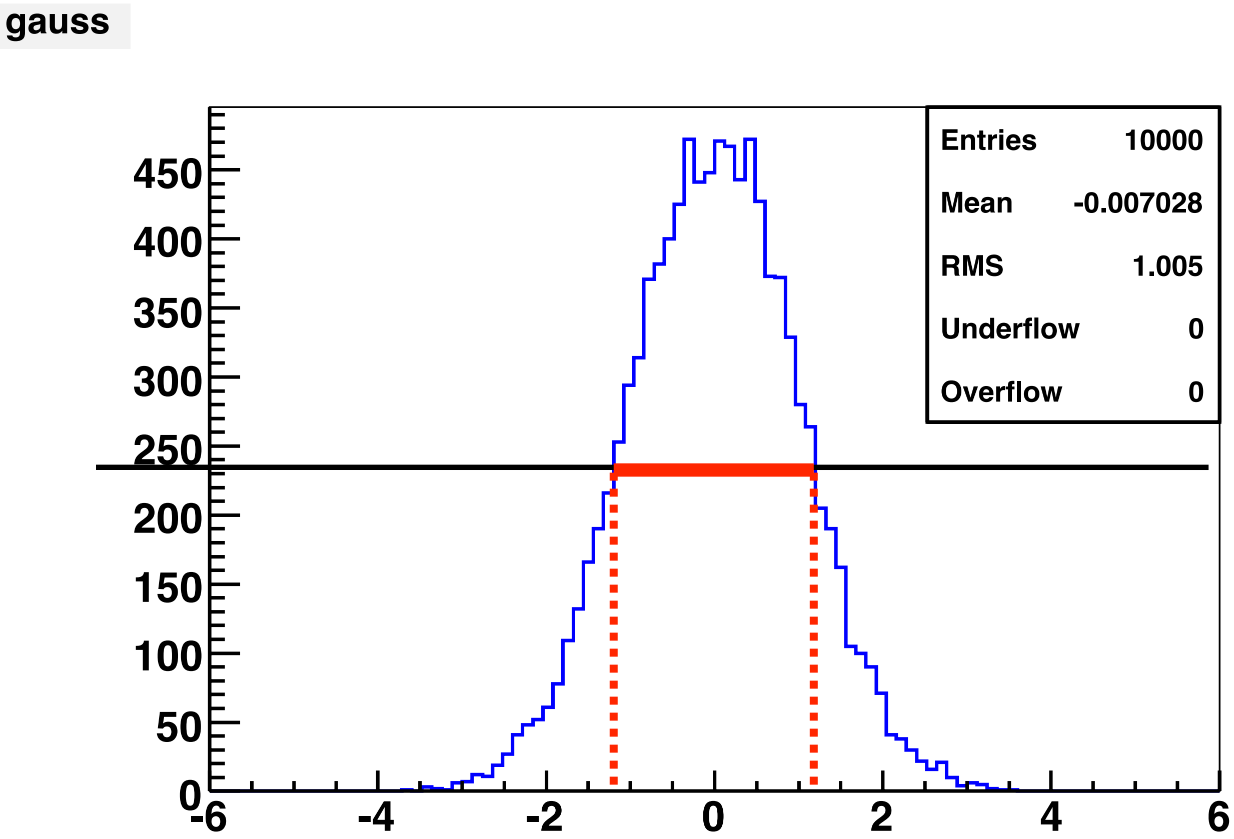
$$\sigma = \sqrt{V}$$

- The Standard Deviation has the same units as the data itself.

- It gives you a “typical” amount by which an individual measurement can be expected to deviate from the mean.

- Usually, a measurement that’s one or two σ away is fine, while 3 σ will raise a few eyebrows. We’ll quantify later what the probabilities for 1, 2, 3 σ deviations are under certain (common) circumstances.

FWHM and standard deviation



- For Gaussian distributions (why these are so important, later):

  FWHM ≈ 2.35σ

- Check histogram on the left:

  σ =RMS = 1.0,

  FWHM= 1.2 – (–1.2) = 2.4

  Close enough.

Covariance

- Consider a data sample where each measurement consists of a pair of numbers: *{(x_1, y_1), (x_2, y_2), ...}*

- The *covariance between x and y* is defined as:

$$\begin{aligned}\operatorname{cov}(x,y) &= \frac{1}{N}\sum_{i=1}^{N}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)\\ &= \overline{xy} - \overline{x}\cdot\overline{y}\end{aligned}$$

- The covariance between two parameters is a quantity that has units; its value depends on the units you chose, difficult to interpret.

Correlation Coefficient

- The correlation coefficient is defined as:

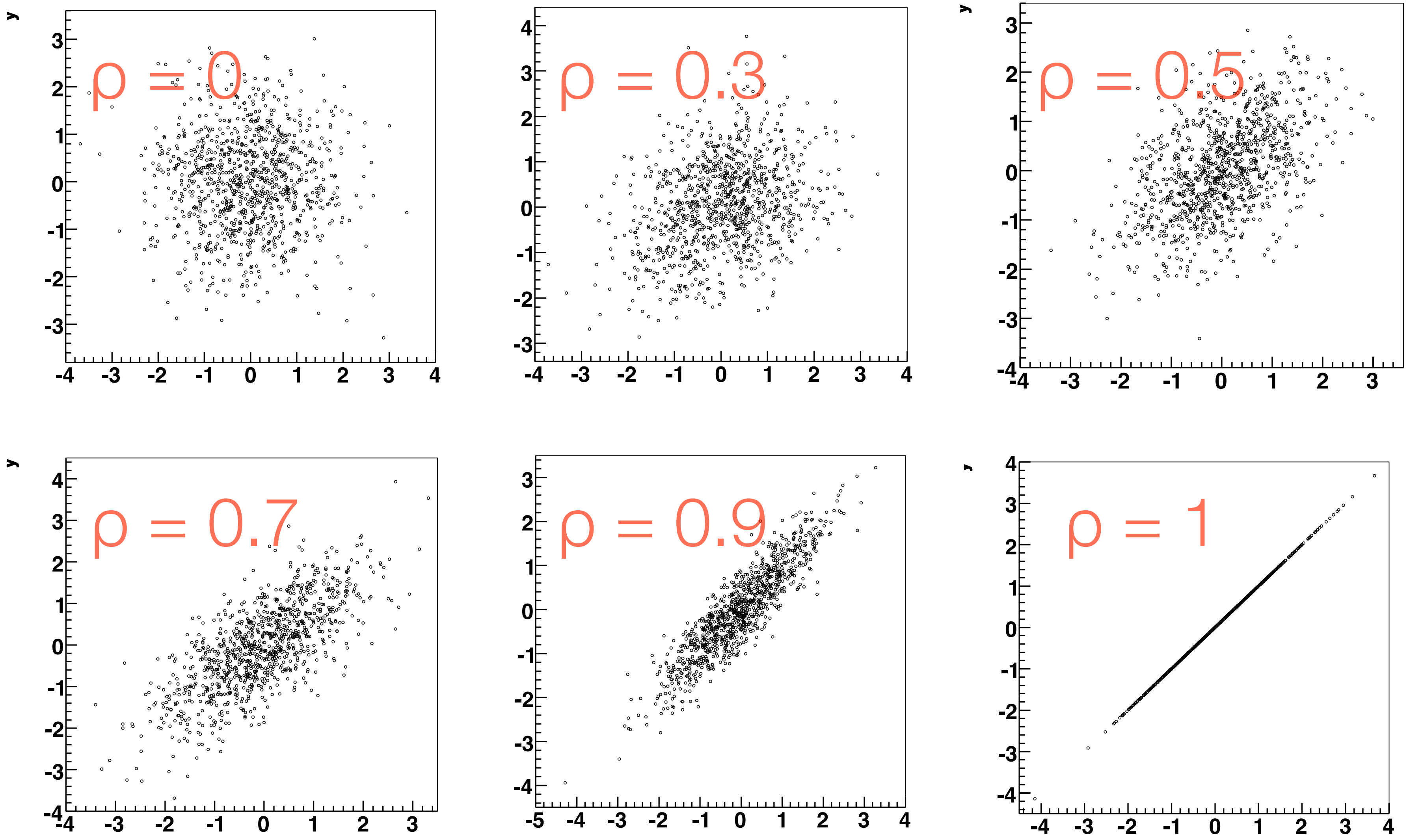
$$\rho_{xy} = \frac{\text{cov}(x,y)}{\sigma_x \cdot \sigma_y}$$

- It has no units and varies between -1 and 1. This provides a measure of how related to quantities are.

- For independent variables, ρ=0 while the correlation coefficient of a parameter with itself (can't get more correlated) is:

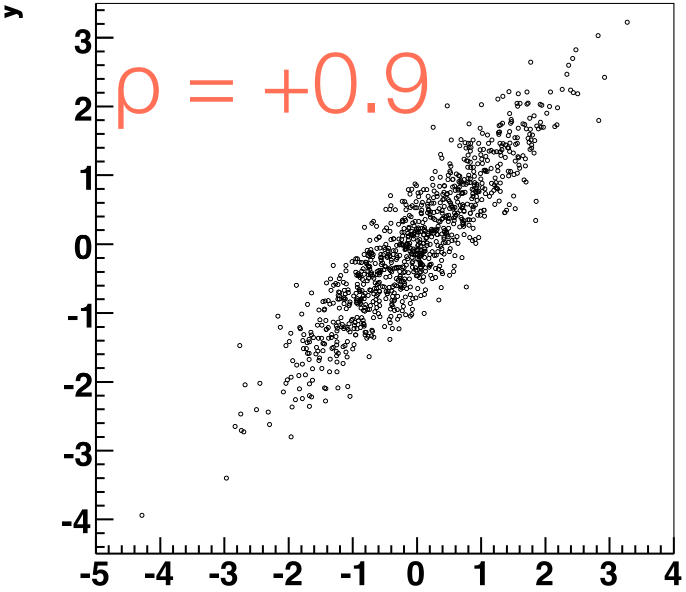
$$\begin{aligned}\rho_{xx} &= \frac{\text{cov}(x,x)}{\sigma_x \cdot \sigma_x} \\ &= \frac{\text{Var}(x)}{\sigma_x^2} = \frac{\sigma_x^2}{\sigma_x^2} = 1\end{aligned}$$

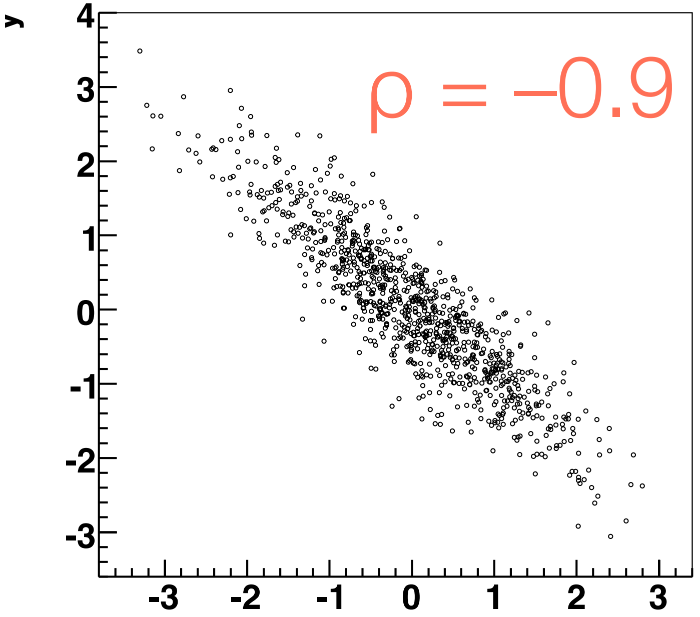
Correlation Coefficient Examples



Correlation Coefficients Examples

- **Correlation coefficients can be positive or negative:**





Make these plots yourself:

https://github.com/JonasRademacker/JupyterNotebooksForTeachingMath/blob/maste

CovarianceAndCorrelation.ipynb

https://tinyurl.com/TeshepStatCode

The Covariance/Error Matrix

- **For N variables, named x^(1)^, ..., x^(N)^**

$$V_{ij} \equiv \text{cov}\left(x^{(i)}, x^{(j)}\right)$$

$$V \equiv \begin{pmatrix} \text{cov}\left(x^{(1)}, x^{(1)}\right) & \text{cov}\left(x^{(1)}, x^{(2)}\right) & \cdots & \text{cov}\left(x^{(1)}, x^{(N)}\right) \\ \text{cov}\left(x^{(2)}, x^{(1)}\right) & \text{cov}\left(x^{(2)}, x^{(2)}\right) & \cdots & \text{cov}\left(x^{(2)}, x^{(N)}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}\left(x^{(N)}, x^{(1)}\right) & \text{cov}\left(x^{(N)}, x^{(2)}\right) & \cdots & \text{cov}\left(x^{(N)}, x^{(N)}\right) \end{pmatrix}$$

- **Symmetric. Diagonal = variances. Off-diagonal: covariances.**

- **Will become very important when we discuss errors and multidimensional parameter transformations.**

The Correlation Matrix

- **Defined equivalently, for N variables $x^{(1)}$, ..., $x^{(N)}$**

$$\rho_{ij} \equiv \frac{\mathrm{cov}\left(x^{(i)}, x^{(j)}\right)}{\sigma_i \sigma_j}$$

$$\rho \equiv \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1N} \\ \rho_{21} & 1 & \cdots & \rho_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{N1} & \rho_{N2} & \cdots & 1 \end{pmatrix}$$

- **symmetric**

- **diagonal = 1**

- **Related to covariance matrix by:**

$$V_{ij} = \rho_{ij}\, \sigma_i \sigma_j$$

Correlation and Causality

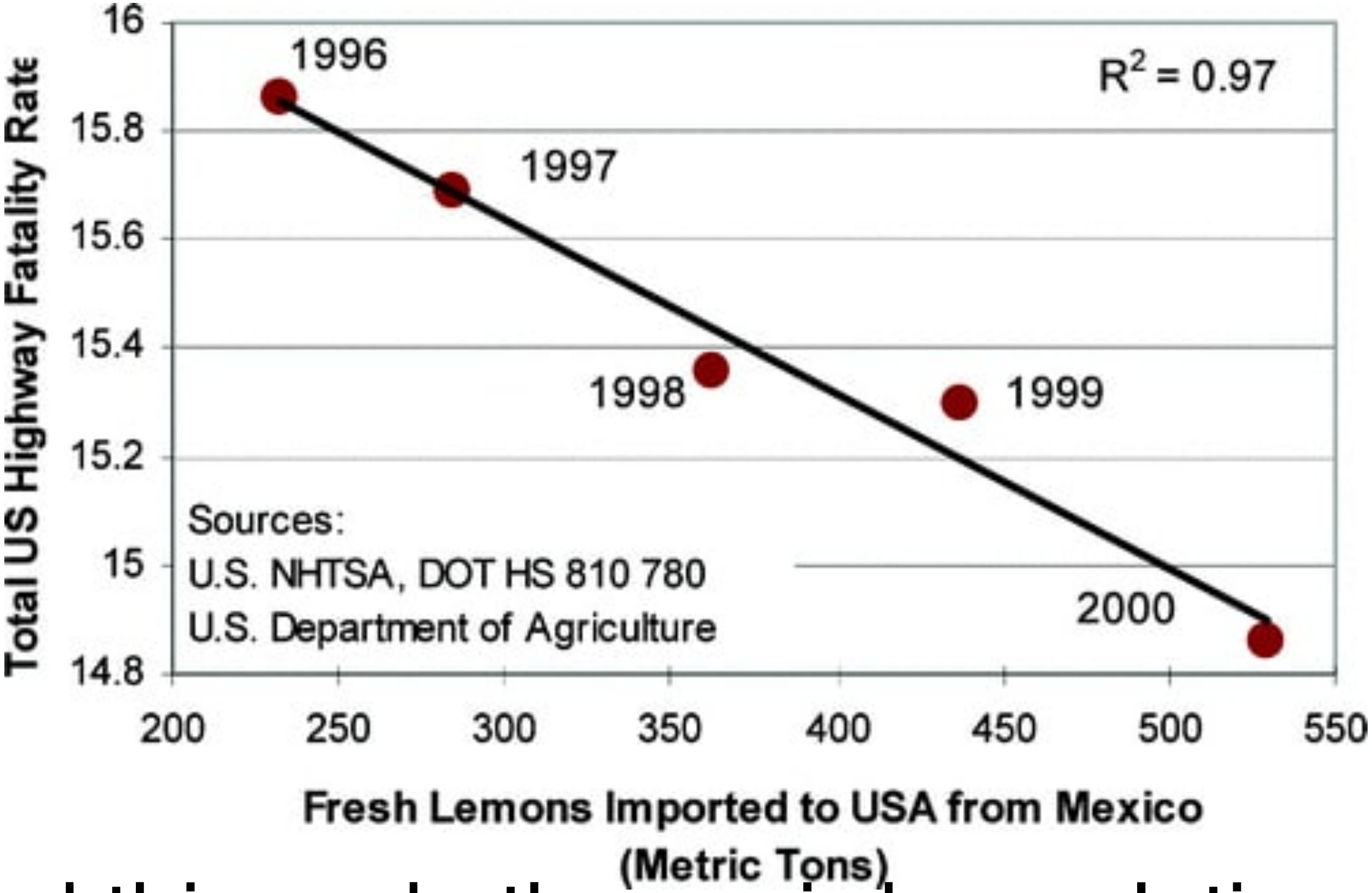
- Among my favourite correlations is this one:

- During doctors' strikes the death-rate tends to go down - in Israel the death-rate went down by 39% in a recent doctors' strike. So there is a positive correlation between life-expectancy and the number of doctors on strike (this phenomenon has been observed in other countries, too). Does this mean that fewer doctors would be good for the nation's health?

- Listen to this BBC programme if you like this sort of thing:

http://news.bbc.co.uk/2/hi/programmes/more_or_less/7408337.stm

Lemons prevent traffic deaths

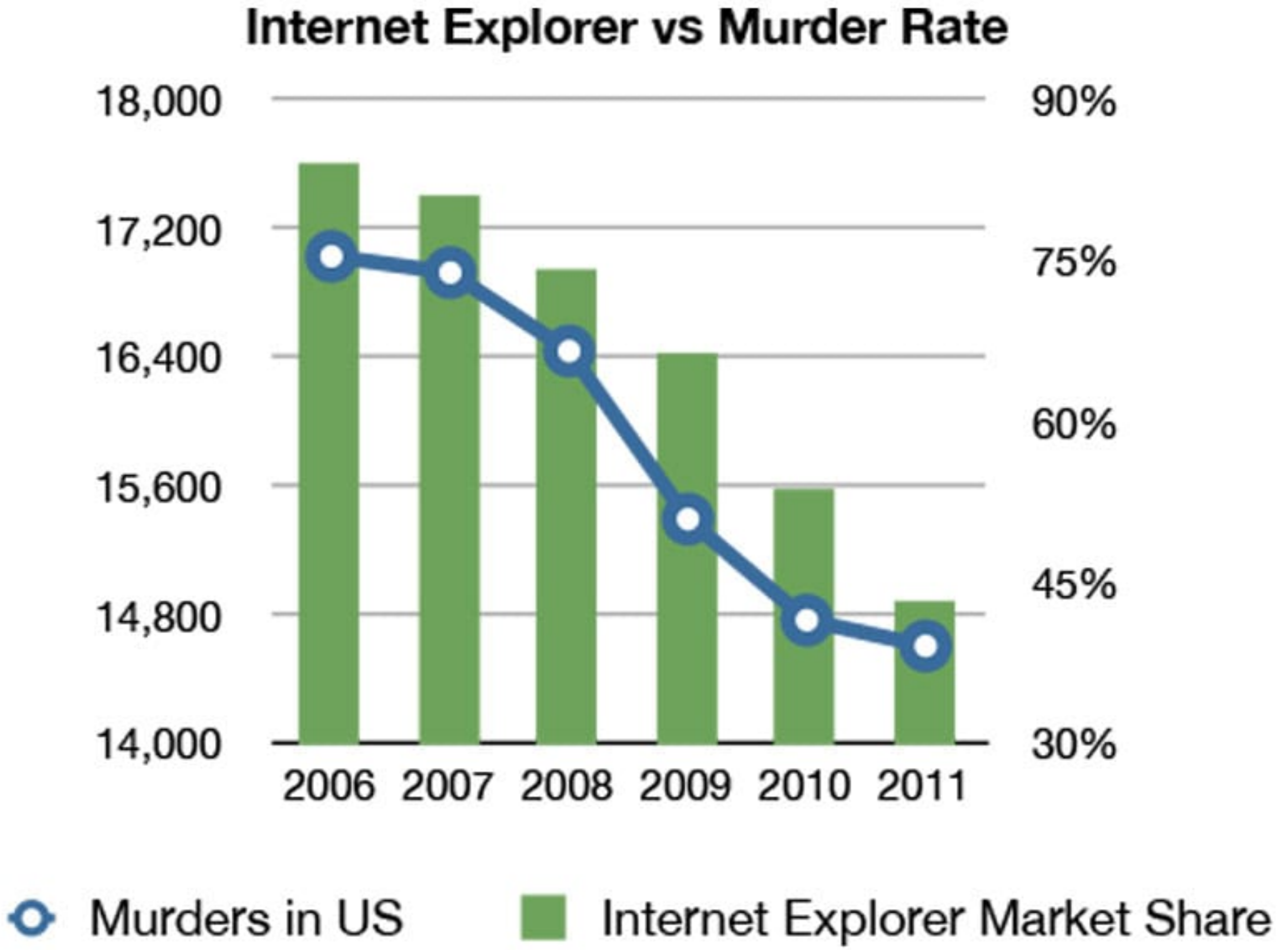


Sources:
U.S. NHTSA, DOT HS 810 780
U.S. Department of Agriculture

find this and other weird correlations at https://www.buzzfeednews.com/article/kjh2110/the-most-bizarre-correlations

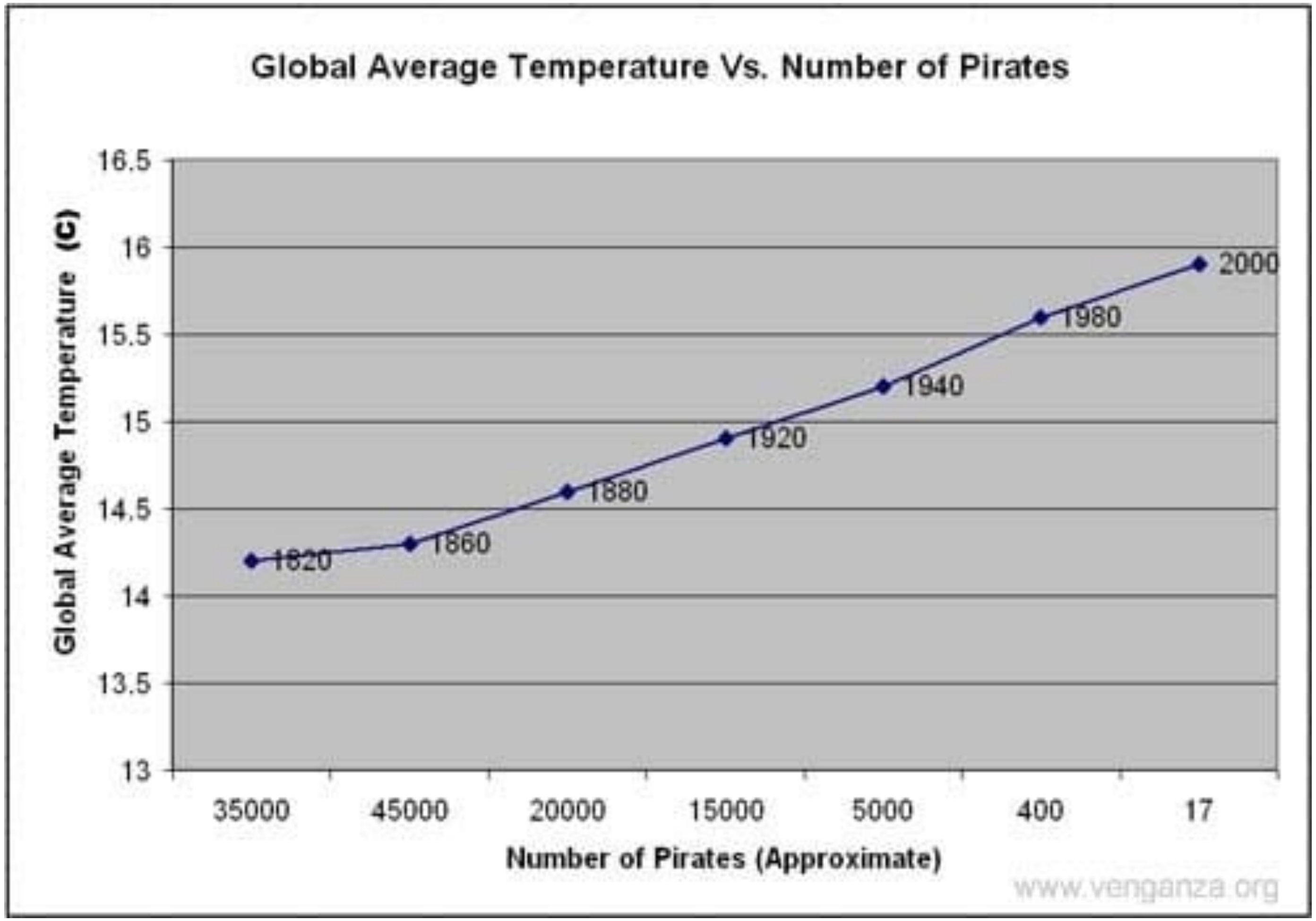
http://pubs.acs.org/doi/abs/10.1021/c

Internet Explorer causes murder



http://gizmodo.com/5977989/internet-explo
murder-rate-will-be-your-favorite-chart-toda

Lack of (Caribbean) pirates causes global warming



http://www.venganza.org/about/open-lette

Correlation and Causality

- Statistics does not tell us if two correlated variables are also connected by causality, i.e. if one causes the other.

- For example there is a strong correlation between rain and wet roads. It is clear that rain causes roads to be wet, and that wet roads do not cause rain. But the statistics won't tell you that.

- There is also a clear correlation between wet roads and the the number of people running around with wet hair. Here neither causes the other, but both are correlated because they have a common cause.

Homework

- **Write down 100 times:**

  “Correlation is not causation”

Summary: Representing Data

- Central value: Usually use arithmetic mean. Nice: Means add up. (i.e. <x + y> = <x> + <y>)

- Width: Use standard deviation. Standard deviations do not add up. Variances do, i.e. V(x+y) = V(x) + V(y) (if variables x and y are uncorrelated).

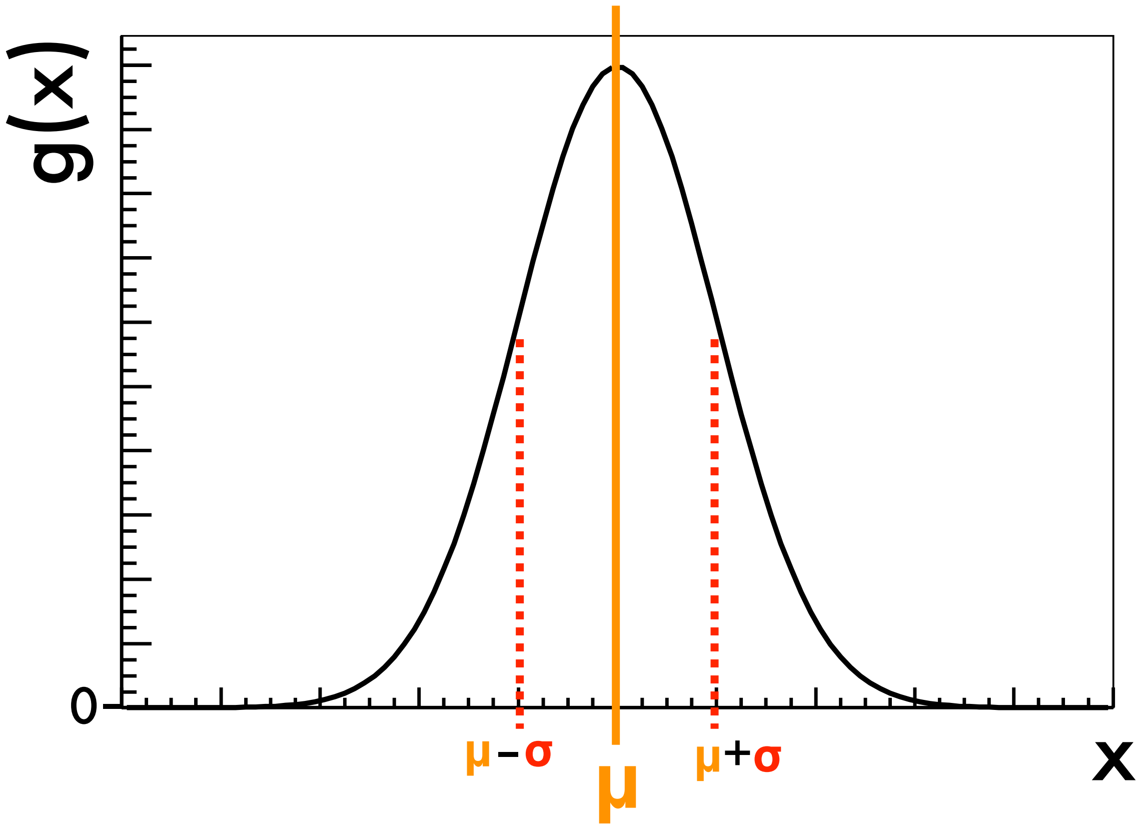
- Multiparameter distributions: Covariance, Correlation.

https://www.youtube.com/watch?v=SSbBvKaM6sk

https://www.youtube.com/watch?v=

We only ever see a slightly blurred picture of nature

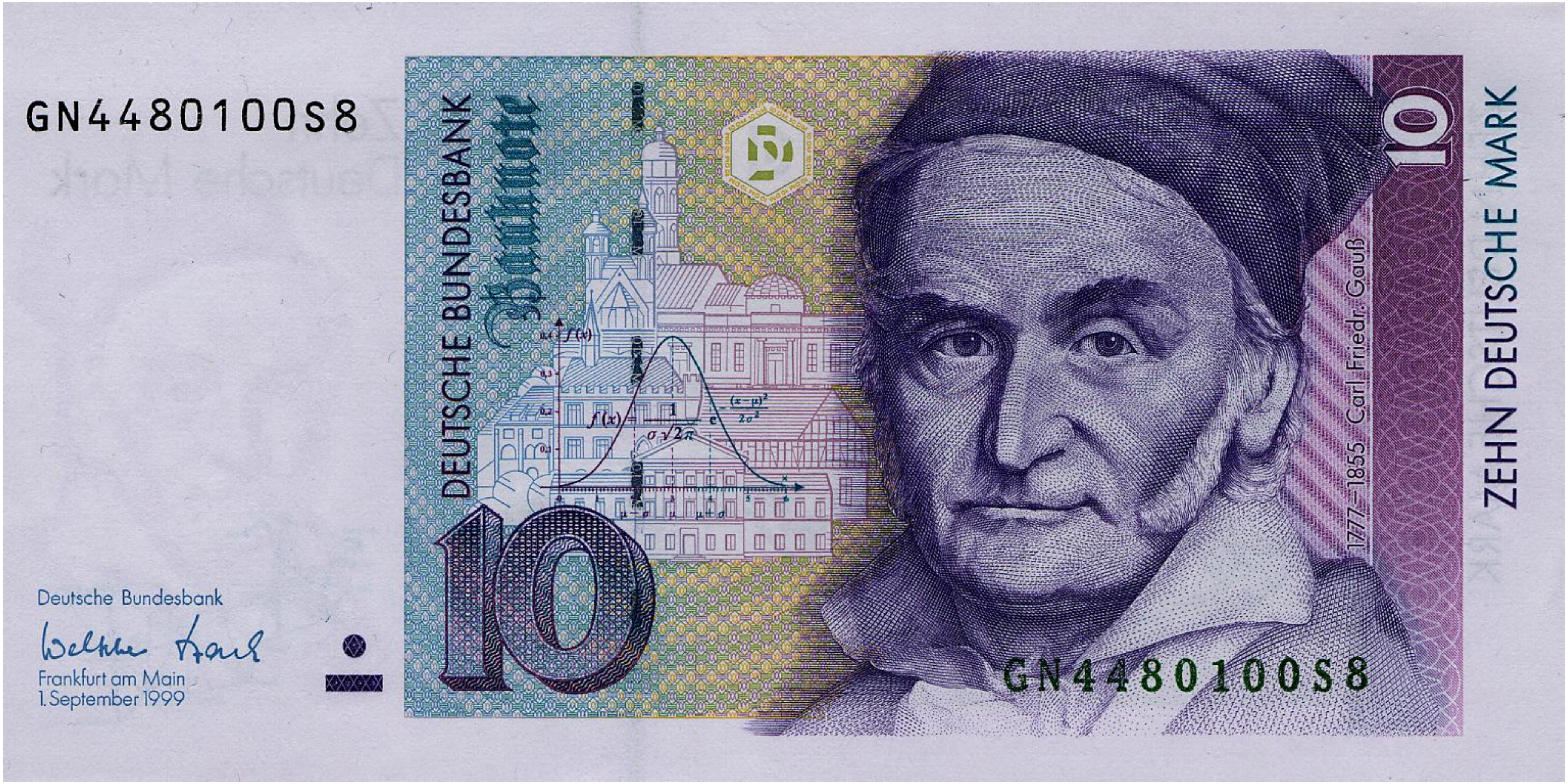
Why the blur is Gaussian



$$g(x;\mu,\sigma) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Gauss & me hanging out in Göttingen

Gauss on old money


GN4480100S8
DEUTSCHE BUNDESBANK
Banknote
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
$\mu-\sigma$ μ $\mu+\sigma$
10
1777–1855 Carl Friedr. Gauß
10
ZEHN DEUTSCHE MARK
Deutsche Bundesbank
Frankfurt am Main
1. September 1999
GN4480100S8

The Central Limit Theorem

- Consider random variable $Y = \sum_i x_i$, where each x_i is taken from a distribution with mean $\langle x_i \rangle$ and variance $V_i = \sigma_i^2$

- Then
    - Y has an expectation value $\langle Y \rangle = \sum_i \langle x_i \rangle$
    - Y has a variance $V_Y = \sum_i V_i$. Equivalently: $\sigma_Y^2 = \sum_i \sigma_i^2$
    - The distribution of *X* becomes Gaussian as *N*→∞.

Variances add up!
(Standard deviations don't)

Roll some Dice, submit results, here

https://tinyurl.com/DiceTESHEP

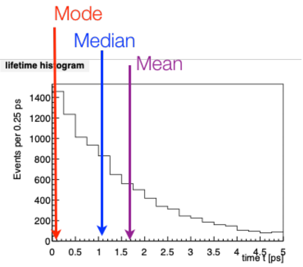
Largest number of entries wins!

Rolling Dice, *predict* results, here

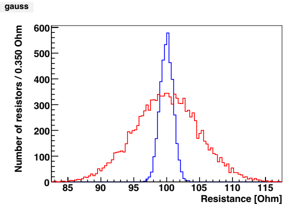
https://tinyurl.com/PredictDiceTESHEP

First (few) correct answers win

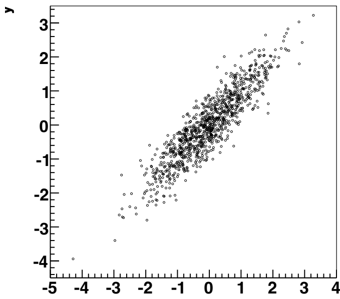
Summary of first statistics lecture



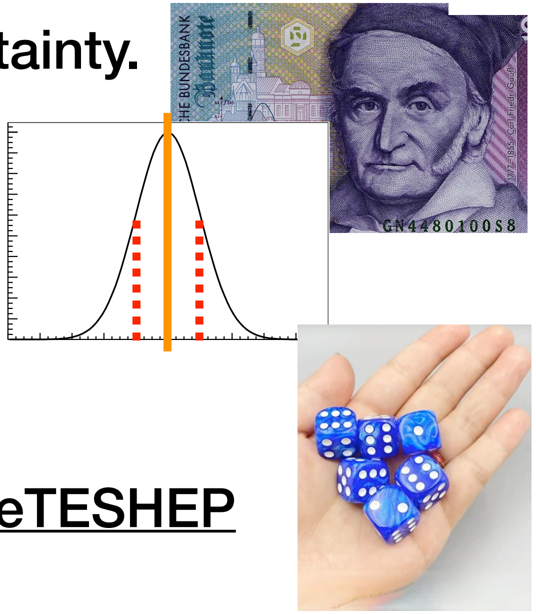
- Averages: Mean, Median, Mode - usually we chose arithmetic mean, but there are use cases for alternatives.



- Width: Standard deviation, Variance, FWHM



- Covariance, correlation (is not causation, but still informative)
- CLT, transforms ignorance to well-defined uncertainty.



- Do your bit for the CLT and win a prize!
  - Roll dice: https://tinyurl.com/DiceTESHEP
  - Predict results: https://tinyurl.com/PredictDiceTESHEP