

Statistical Methods in Particle Physics

Introduction

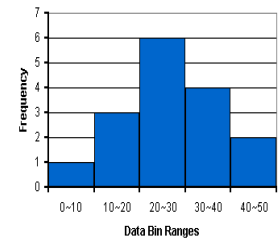
October 15, 2012

Silvia Masciocchi, GSI Darmstadt
s.masciocchi@gsi.de

Niklaus Berger, University of Heidelberg
nberger@physi.uni-heidelberg.de

Winter Semester 2012 / 13

Information about the course



Fakultät für Physik und Astronomie

Master of Science Physics

Specialised Courses

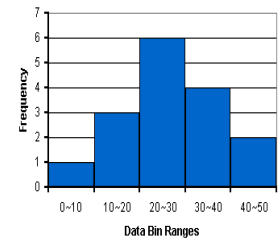
Particle Physics [MVP] - - - 1

Graduate Education [G] (HGSFP, IMPRS-HD, IMPRS-PTFS, IMPRS-QD)

Particle Physics and Cosmology [GP] - - - 2

Day	Time	Frequency	Room	Teacher
Monday	16:15 - 18:00	weekly	SR 2.401, KIP Im Neuenheimer Feld 227	Silvia Masciocchi
Monday	18:00 – 19:00	weekly	CIP Pool, PI Im Neuenheimer Feld 226	Niklaus Berger

Information about the course



How to reach us:

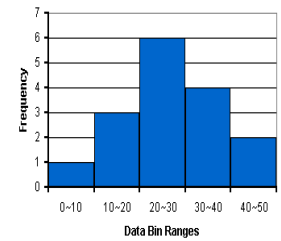
- Silvia Masciocchi
GSI Helmholtzzentrum für Schwerionenforschung, Darmstadt
s.masciocchi@gsi.de Tel. 06151 - 71 1489 KWB 5.27
- Niklaus Berger
Physikalisches Institut, Heidelberg room 2.404
nberger@physi.uni-heidelberg.de Tel. 06221 - 54 9342

The course includes:

- Lectures
- Exercises = computer course
- Homeworks

} **4 Credit Points**

The web page



<http://www.physi.uni-heidelberg.de/~nberger/teaching/ws12/statistics/statistics.php>

Statistical Methods in Particle Physics WS 2012/2013

[S. Masciocchi \(lectures\)](#) / [N. Berger \(exercises\)](#)

Lectures every Monday 16:15 - 18:00, starting October 15th at [Seminar room 2.401](#), KIP, Im Neuenheimer Feld 227

Exercises every Monday 18:00 - 19:00, starting October 15th at CIP Pool, PI, Im Neuenheimer Feld 226

Please sign up at [the electronic tutorial administration](#)

Lectures

15.10.2011 [Lecture 1](#) Introduction: Aims of the course, distributions and their properties, histograms

Exercises

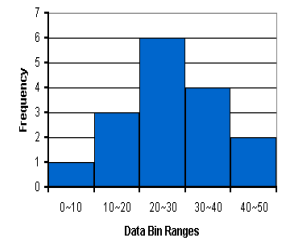
The exercises will be held on the CIP pool computers and involve writing scripts and programs in C++ using the root data analysis framework, putting to work the concepts taught in the lecture. For help and documentation with the tools, see [here](#).

15.10.2011 [Root tutorial \(Exercise 0\)](#)
[Exercise 1](#) Introduction to the root framework, histograms

Thanks Nik !!!

Slides of lectures and exercises will be uploaded

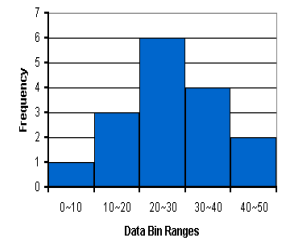
Why do we need statistics in physics?



- **Experimental measurements** are only **SAMPLES** of the reality, they can never represent the entire set of possibilities
 - they are affected by uncertainties
 - results can be expressed as probabilities
- **Theoretical calculations** are mostly **APPROXIMATIONS** limited by finite resources to do the calculations or by imprecise input parameters
 - are also affected by uncertainties
 - predictions can also be expressed in terms of probability

understand the role of **uncertainty and **probability** in relating data and theory !!**

Statistical tools

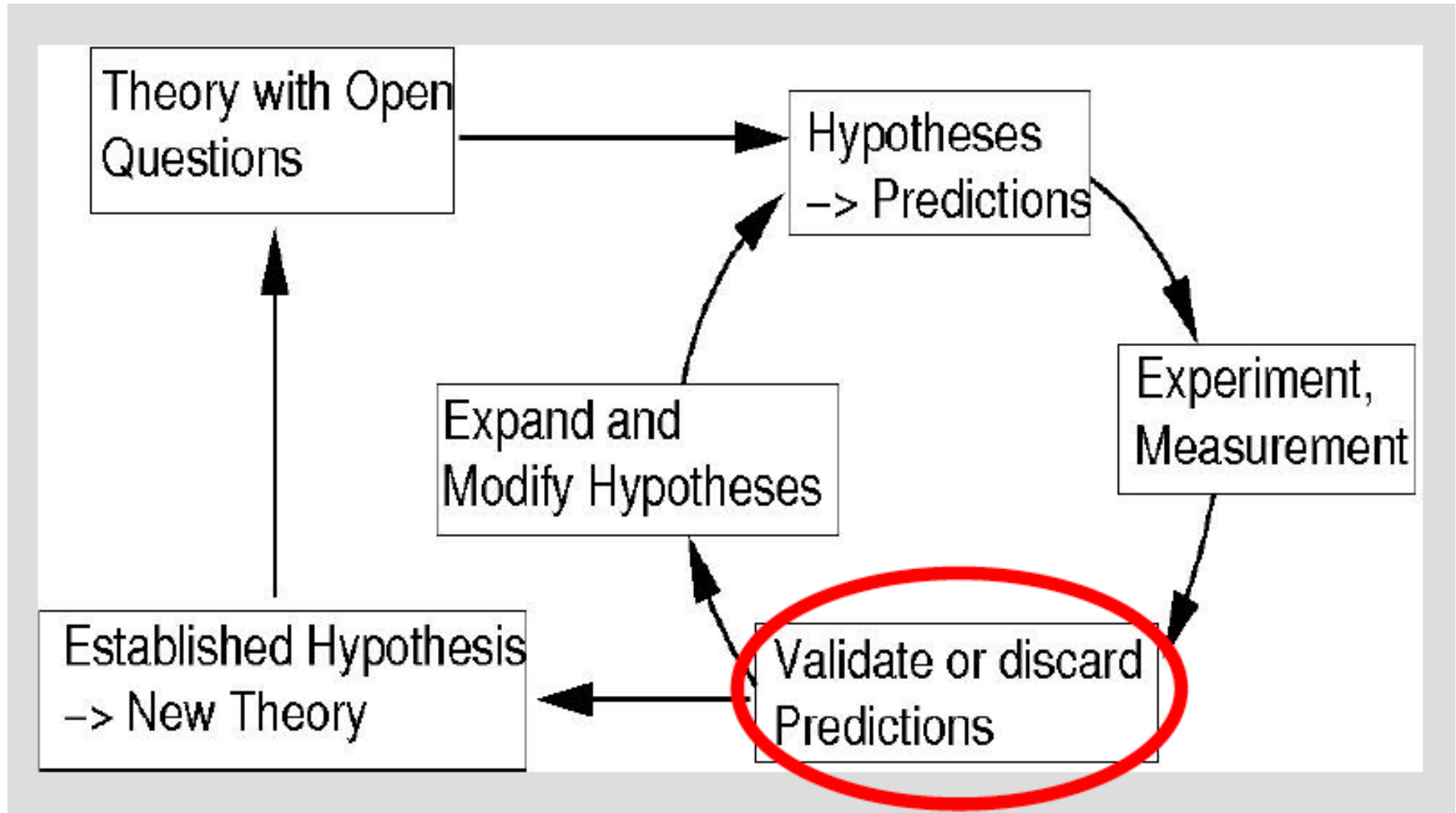
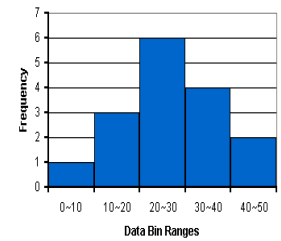


The analysis of experimental data requires **statistical tools** for example:

- Assign uncertainty / error to measurements
- Error propagation
- Appropriate data reduction and representation
- Parametrization of distributions, fitting procedures
- Go from the measurements to the extraction of physical quantities

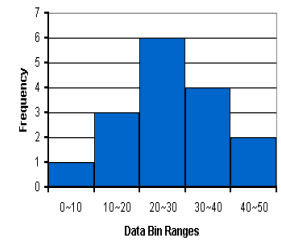
→ In this course: learn the tools, and practice them!!

Physics methodology



STATISTICAL METHODS ARE CRUCIAL !!!

One example



The Standard Model of particle physics

Elementary particles

Heavy Flavours:

Charm

Beauty

Three Generations of Matter (Fermions)

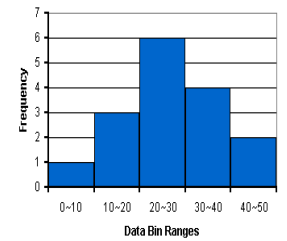
	I	II	III	
mass →	2,4 MeV	1,27 GeV	171,2 GeV	0
charge →	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	0
spin →	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1
name →	u up	c charm	t top	γ photon
	4,8 MeV	104 MeV	4,2 GeV	0
	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$	0
	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1
	d down	s strange	b bottom	g gluon
	<2,2 eV	<0,17 MeV	<15,5 MeV	91,2 GeV
	0	0	0	0
	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1
	ν_e electron neutrino	ν_μ muon neutrino	ν_τ tau neutrino	Z⁰ Z boson
	0,511 MeV	105,7 MeV	1,777 GeV	80,4 GeV
	-1	-1	-1	± 1
	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1
	e electron	μ muon	τ tau	W[±] W boson

Quarks

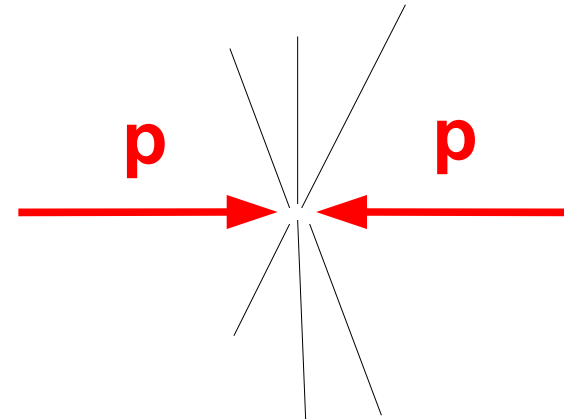
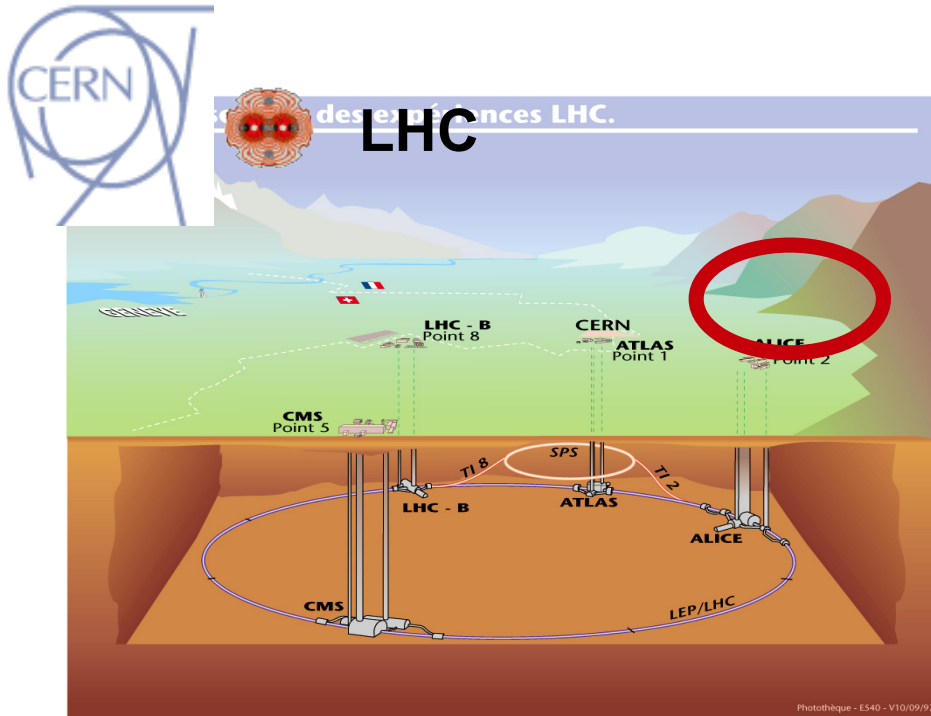
Leptons

Gauge Bosons

One example

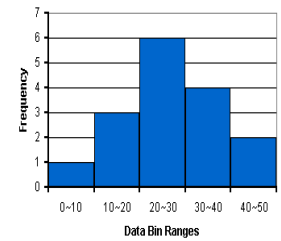


- I am interested in particles (hadrons) which contain heavy flavours (charm, beauty)
- I want to know how many of those are produced in collisions of protons (p-p) at LHC, at center-of-mass energy of 7 TeV



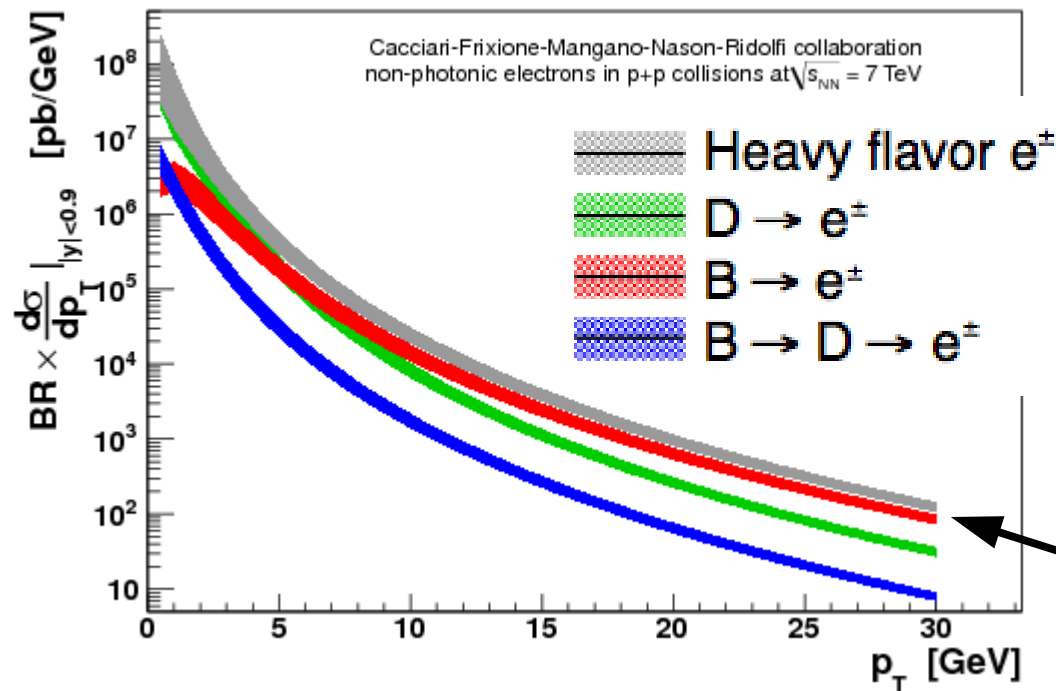
e.g. charm quark
→ D meson
→ **electron** + more
(decay)

One example: from theory ...



- Theory prediction:
FONLL (**F**ixed **O**rders plus **N**ext-to-**L**eading **L**ogarithms)

Predicts distribution of electrons from hadrons with charm and beauty

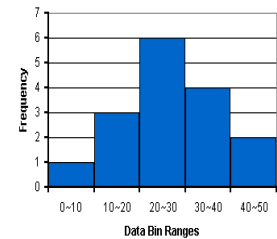


BUT:

**WITH
UNCERTAINTIES !!!**

“band” of possible values !!!

One example: ... to measurement

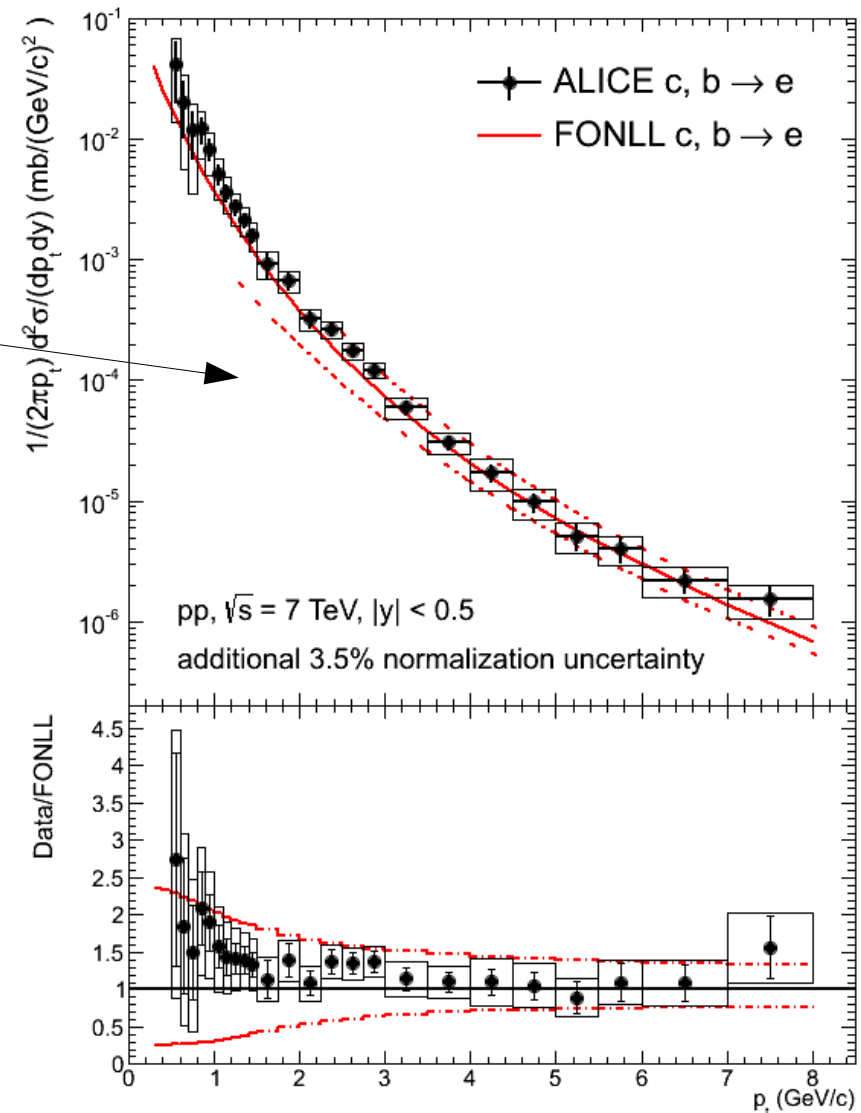


We measure that distribution of electrons in ALICE

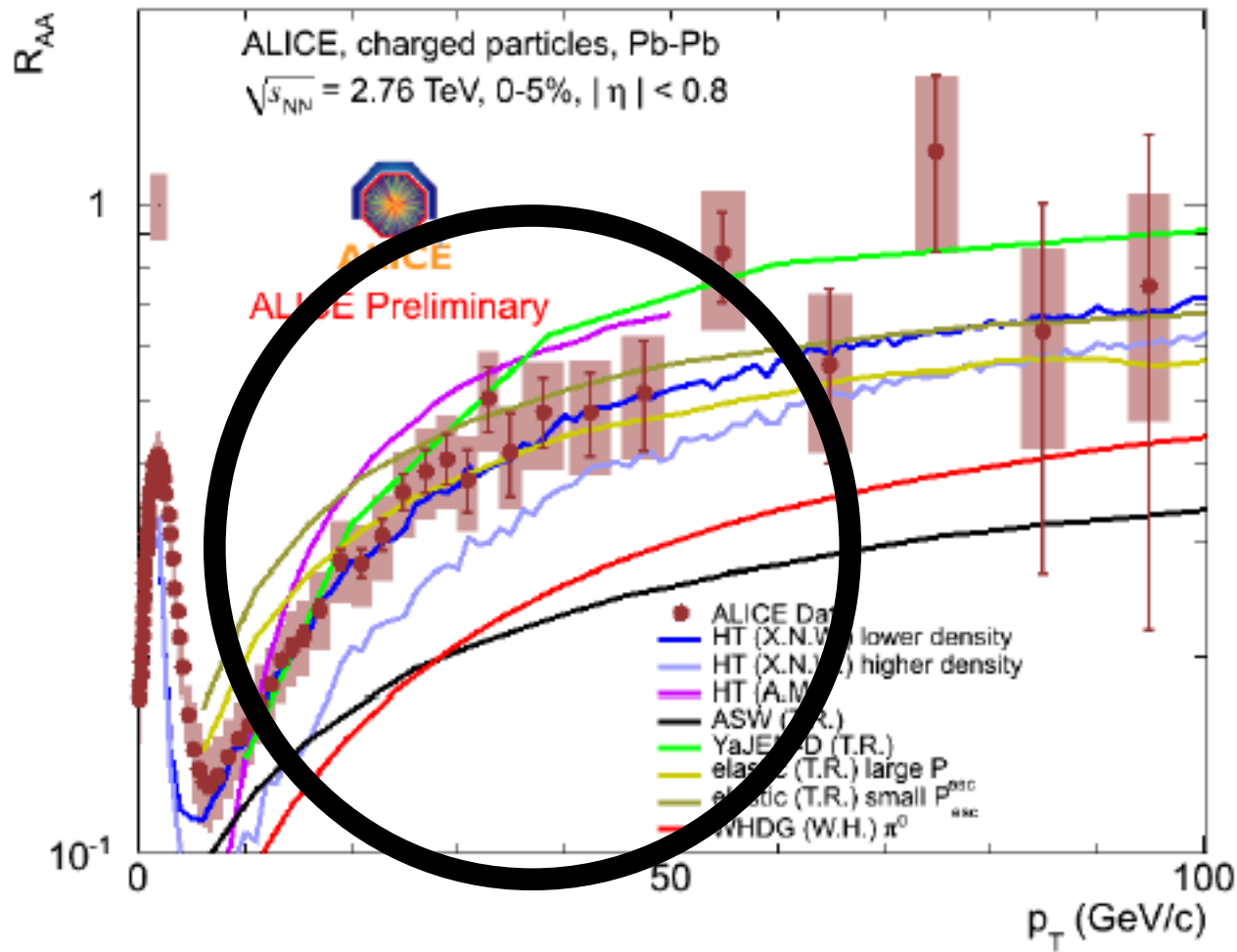
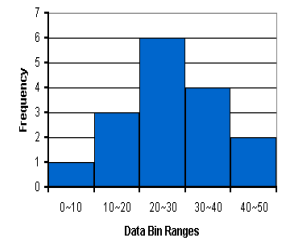
- **measurement UNCERTAINTIES !!!**

And compare measurement with theory (here: ratio of the two)

Yes, they agree



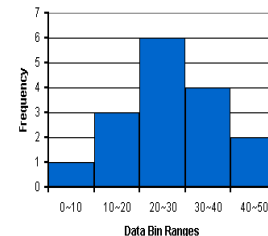
Sometimes they do NOT agree ...



**Measurements can confirm predictions or not
→ theories evolve, models are taken or discarded**

Statistical methods are mandatory

And viceversa



In the comparison between:

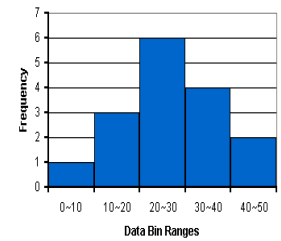
- The state of the art of a measurement
and
- The current theory predictions

we can have that, for example:

- A measurement can be so imprecise that it cannot discriminate between different predictions
- Two measurements (two experiments) are incompatible but also imprecise such that there cannot be a conclusion

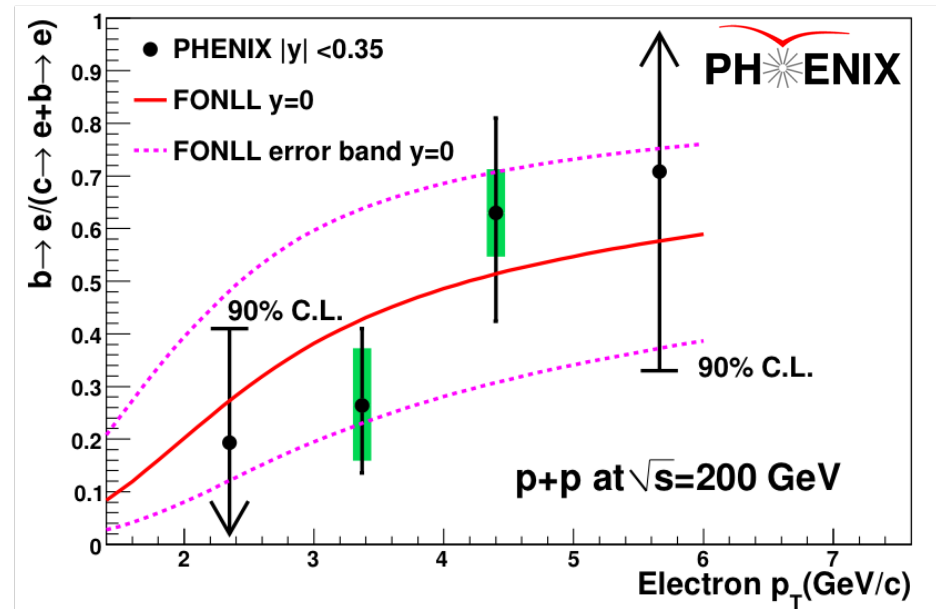
This can determine future experiments, the design of the next generation detectors, etc ...

Too limited precision



Back to heavy flavours:

Ratio $\frac{\text{“beauty”}}{\text{“beauty+charm”}}$
in PHENIX @ RHIC

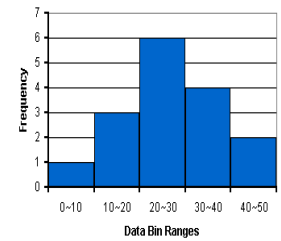


At RHIC, charm and beauty cannot be really separated →

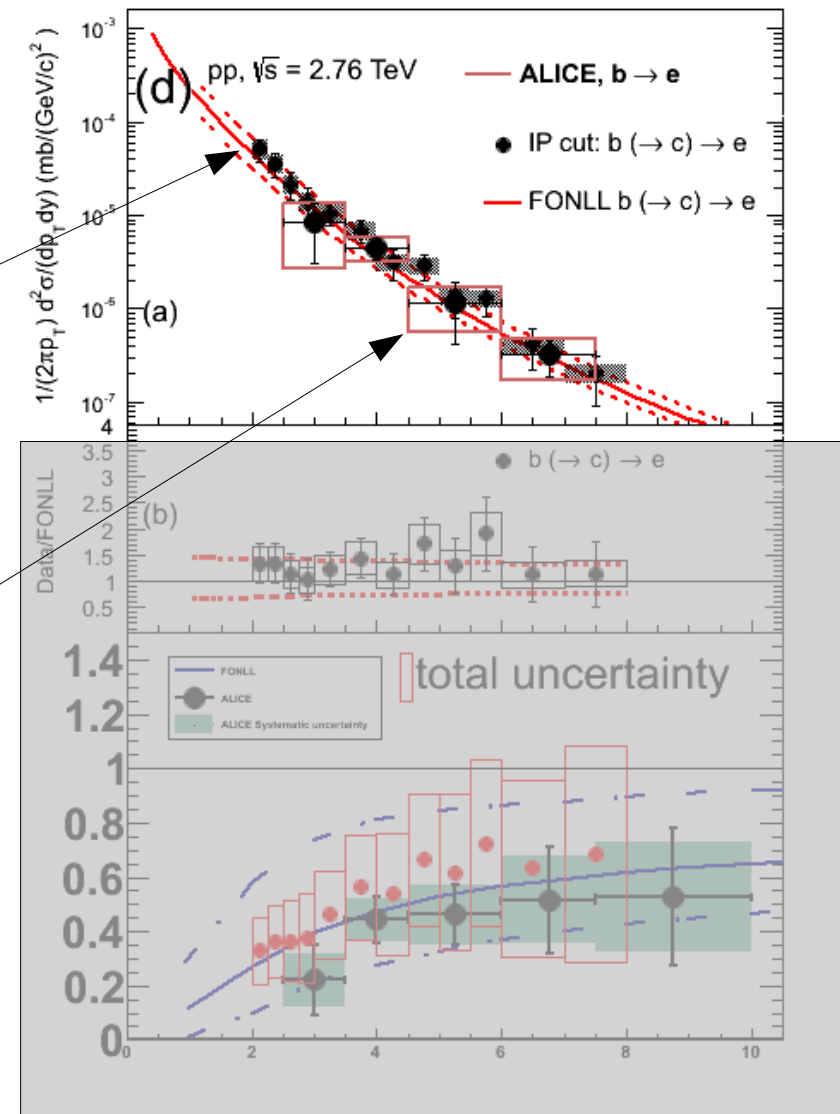
Results affected by extremely large uncertainty → not decisive

This influenced the design of ALICE @ LHC, particularly its vertex detector !!!

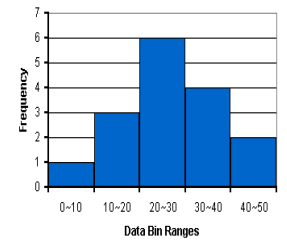
... and how to address that



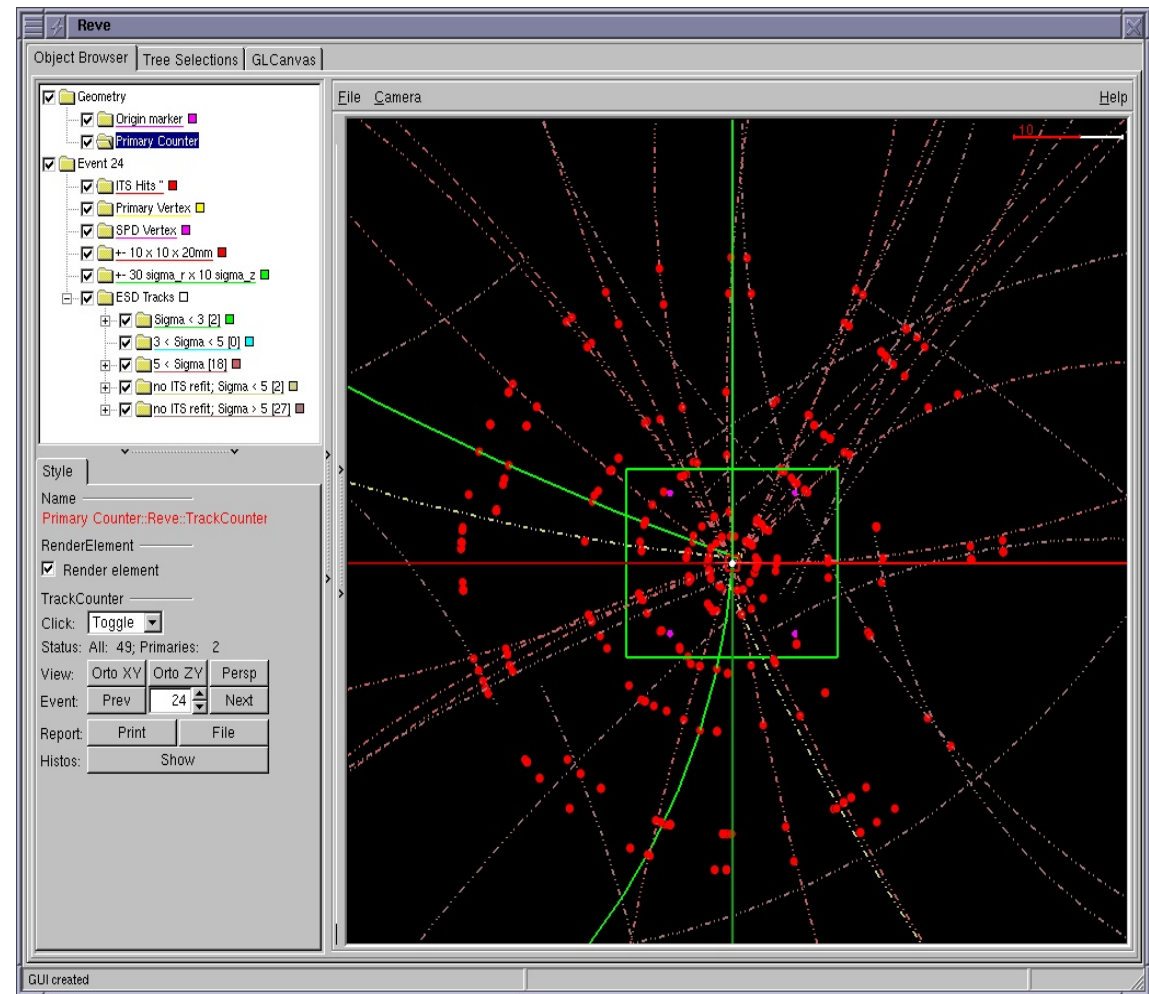
- ALICE has a silicon vertex detector which provides very high spatial resolution
- With such resolution we can measure electrons from beauty hadron decays
- The precision of this method beats by large the correlation method
- This makes POSSIBLE at all the measurement of the production cross section



Down to everyday life

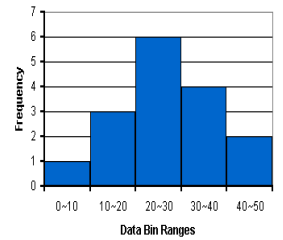


- Particle trajectories in an experimental apparatus
- Particle in detectors leaves a “signal”
- → Points measured with ERRORS
- Reconstruction of “tracks”
- Fitting procedures



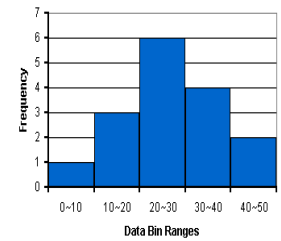
ALICE event display

Lecture program



- Basic concepts and definitions
- Random numbers
- Characteristics of distributions
- Important distributions
- Error propagation
- Fitting procedures
- Estimators:
 - Maximum likelihood
 - Least square method
- Confidence level and limits
- Hypothesis tests
- Limits

The cheating baker



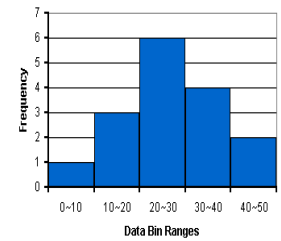
Once upon a time, in a holiday resort the landlord L. ran a profitable Bed&Breakfast, and every morning bought 30 rolls for breakfast. By law the mass of a single roll was required to be 75 g.

One day the owner of the bakery changed, and L. suspected that the new baker B. might be cheating. So he decided to check the mass of what he bought, using a kitchen scale with a resolution of 1 g.

After one month he had collected a fair amount of data:

73 79 72 62 67 60 60 67 78 68 66 75 76 73 75 64 70 69 73 59 70 73 64 72 64 69
69 71 69 71 77 69 72 71 67 72 63 66 68 76 71 76 68 71 63 65 65 66 73 73 73 67
70 65 71 69 78 67 65 69 71 71 72 73 72 69 66 66 70 60 72 62 53 65 74 65 68 69
67 75 64 76 72 76 78 67 67 67 69 79 71 67 71 68 71 65 66 65 78 76 71 70 67 65
67 64 73 67 74 79 74 71 73 67 66 76 68 74 76 65 77 67 71 67 71 77 63 66 70 62
68 74 67 67 67 77 65 68 79 72 71 77 68 70 73 67 81 70 74 71 79 62 67 63 68 76
73 81 76 73 68 72 76 61 69 73 71 80 68 70 62 76 58 68 68 64 68 78 69 65 70 70
64 75 73 72 60 86 68 68 64 60 68 71 70 75 70 67 69 67 73 65 66 71 70 70 73 66
72 71 71 64 76 75 72 72 71 72 72 71 75 68 73 70 64 76 72 75 79 70 64 70 67 70
75 70 83 69 61 70 66 69 71 72 70 76 73 62 71 60 73 74 70 68 68 70 78 71 69 71
73 73 75 65 71 67 60 70 77 71 74 64 74 73 60 77 73 70 69 66 70 78 69 75 66 71
75 75 74 69 74 70 75 77 75 66 72 68 72 61 75 65 69 68 65 73 82 67 75 67 80 71
79 72 71 68 73 70 67 75 74 69 63 63 72 70 73 63 70 70 59 78 76 66 72 79 65 71
76 72 69 69 73 70 77 73 83 66 68 67 69 73 76 65 71 70 71 65 78 71 67 70 72 75
67 79 72 64 62 79 68 70 61 65 68 71 73 60 60 68 71 74 75 69 73 70 68 ...

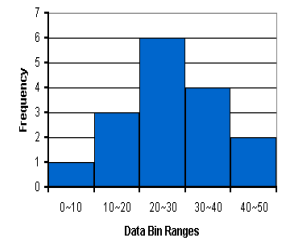
Data reduction



- The raw list of numbers is not very useful!
→ we need some kind of **data reduction** !
- Assume that all measurements are equivalent:
 - The sequence of individual data does not matter
 - All relevant information is contained in the number of counts per reading

count [50]= 0	count [60]= 20	count [70]= 85	count [80]= 9
count [51]= 0	count [61]= 11	count [71]= 81	count [81]= 7
count [52]= 0	count [62]= 20	count [72]= 61	count [82]= 3
count [53]= 0	count [63]= 21	count [73]= 65	count [83]= 5
count [54]= 0	count [64]= 31	count [74]= 54	count [84]= 0
count [55]= 0	count [65]= 48	count [75]= 43	count [85]= 0
count [56]= 2	count [66]= 42	count [76]= 33	count [86]= 1
count [57]= 1	count [67]= 70	count [77]= 23	count [87]= 0
count [58]= 3	count [68]= 68	count [78]= 21	count [88]= 0
count [59]= 6	count [69]= 74	count [79]= 20	count [89]= 1

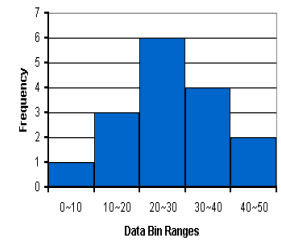
Data reduction



count [50]= 0	count [60]= 20	count [70]= 85	count [80]= 9
count [51]= 0	count [61]= 11	count [71]= 81	count [81]= 7
count [52]= 0	count [62]= 20	count [72]= 61	count [82]= 3
count [53]= 0	count [63]= 21	count [73]= 65	count [83]= 5
count [54]= 0	count [64]= 31	count [74]= 54	count [84]= 0
count [55]= 0	count [65]= 48	count [75]= 43	count [85]= 0
count [56]= 2	count [66]= 42	count [76]= 33	count [86]= 1
count [57]= 1	count [67]= 70	count [77]= 23	count [87]= 0
count [58]= 3	count [68]= 68	count [78]= 21	count [88]= 0
count [59]= 6	count [69]= 74	count [79]= 20	count [89]= 1

- Much improved presentation of the collected information
- The numbers above cover the entire data set
- **Most of the measurements are lower than 75 g**
- Improve representation, with visual one !

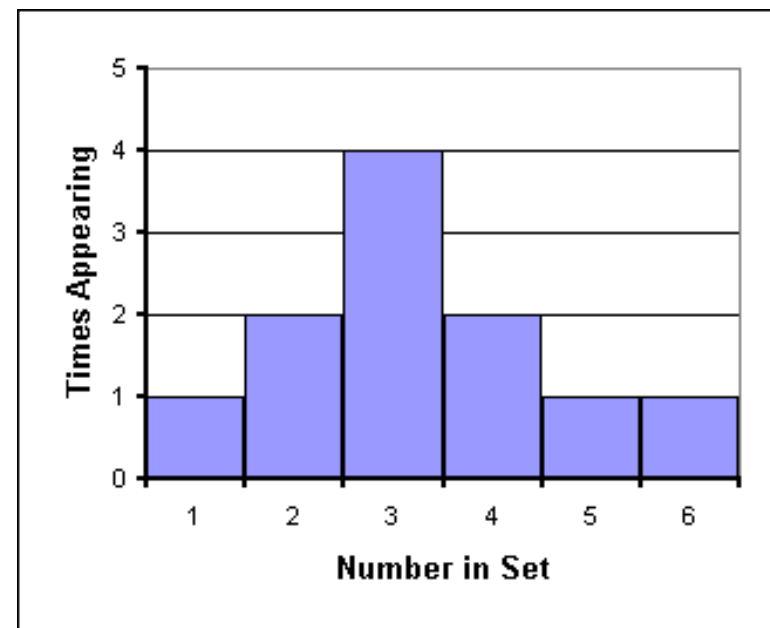
What is a histogram



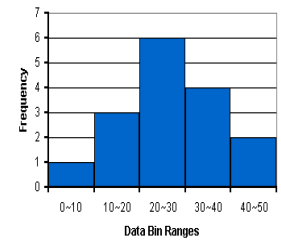
A histogram is "a representation of a frequency distribution by means of rectangles whose **widths** represent class intervals and whose **areas** are proportional to the corresponding frequencies."

Webster's Dictionary

Also called bar-chart



What is a histogram

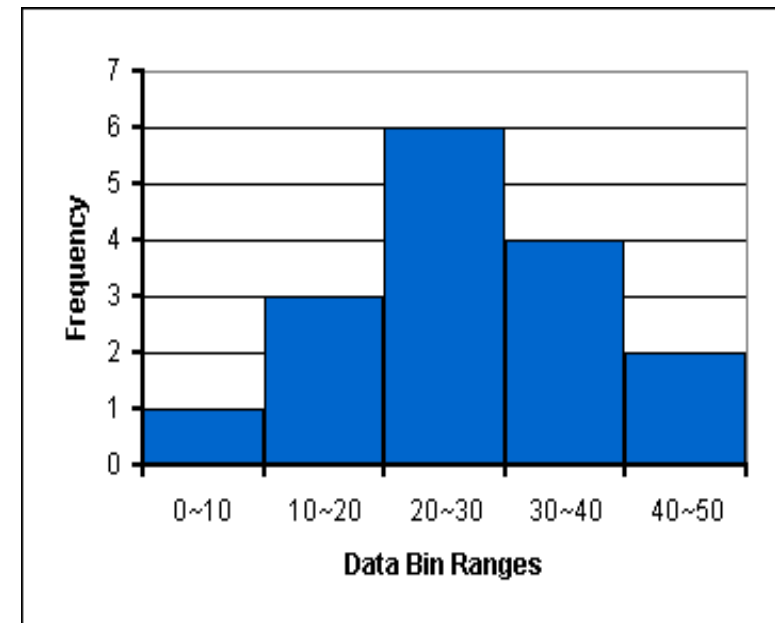


- Horizontal axis represents the quantity of interest, a variable
- Define **bins** for the possible values of the variable (ranges)
- Count the entries in each bin
- Draw a bar of that size

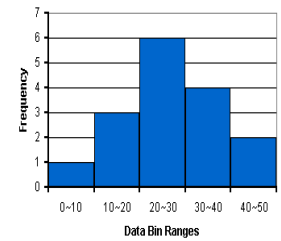
Data Range	Frequency
0-10	1
10-20	3
20-30	6
30-40	4
40-50	2

The visualization gives an impression of the distribution:

- Peak
- Center of the distribution
- Width
- Shape

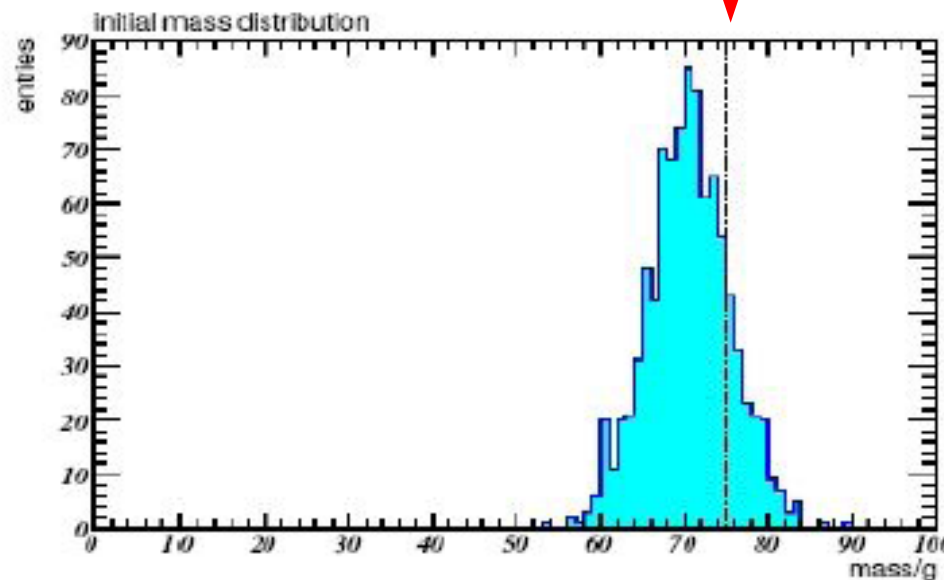


Histogram of rolls



- We already grouped the individual measurements in counts per reading of weight
- Bin: 1 g

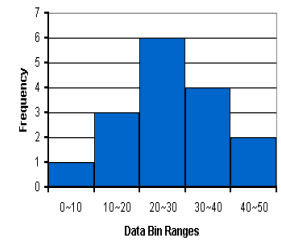
Prescribed weight: 75 g



Symmetric
distribution

The baker B is definitely cheating, his rolls are too light and show a lack of dough

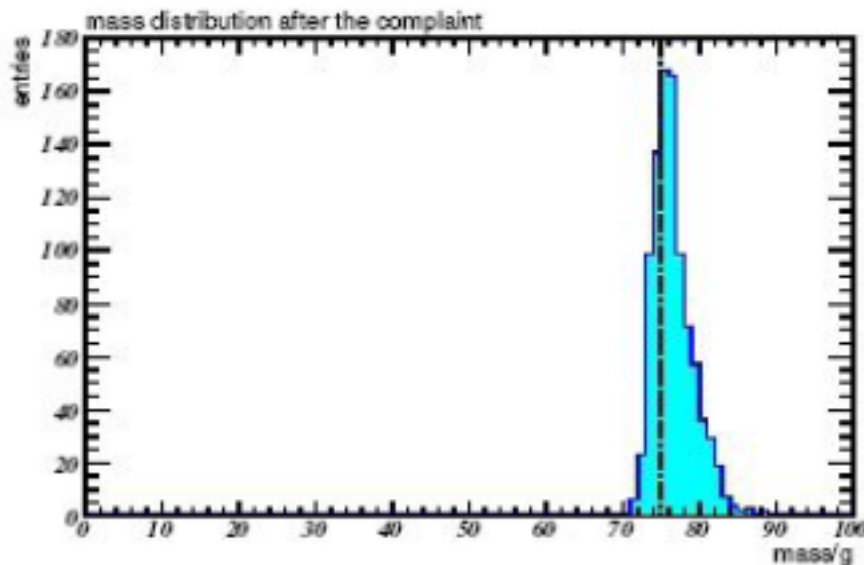
Keeping an eye on the baker



As a consequence of his findings, L. complained to B.

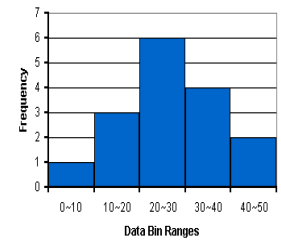
B. apologized and claimed that the low weight of the rolls was an accident which will be corrected in the future. L. however continues to monitor the quality delivered by the baker. One month later, B. inquired again about his products, asking whether now everything is alright.

L. acknowledged that the weight of the rolls now matched his expectations, but he also voiced the opinion that B. was cheating ...



What do you notice in this distribution ??

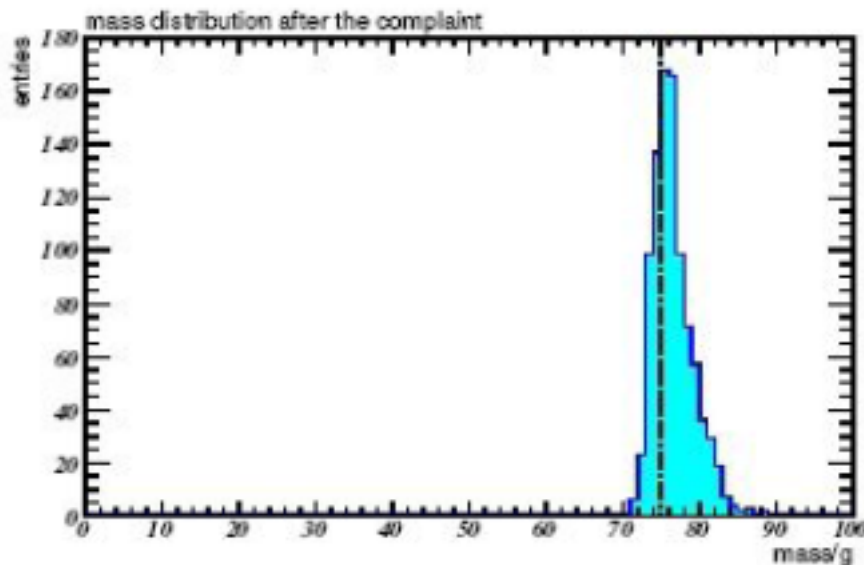
Keeping an eye on the baker



As a consequence of his findings, L. complained to B.

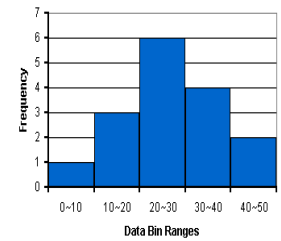
B. apologized and claimed that the low weight of the rolls was an accident which will be corrected in the future. L. however continues to monitor the quality delivered by the baker. One month later, B. inquired again about his products, asking whether now everything is alright.

L. acknowledged that the weight of the rolls now matched his expectations, but he also voiced the opinion that B. was cheating ...



B. simply selects the heaviest rolls for L. !!!

One more histogram



Science 6 July 2007:
Vol. 317 no. 5834 p. 82
DOI: 10.1126/science.1139940

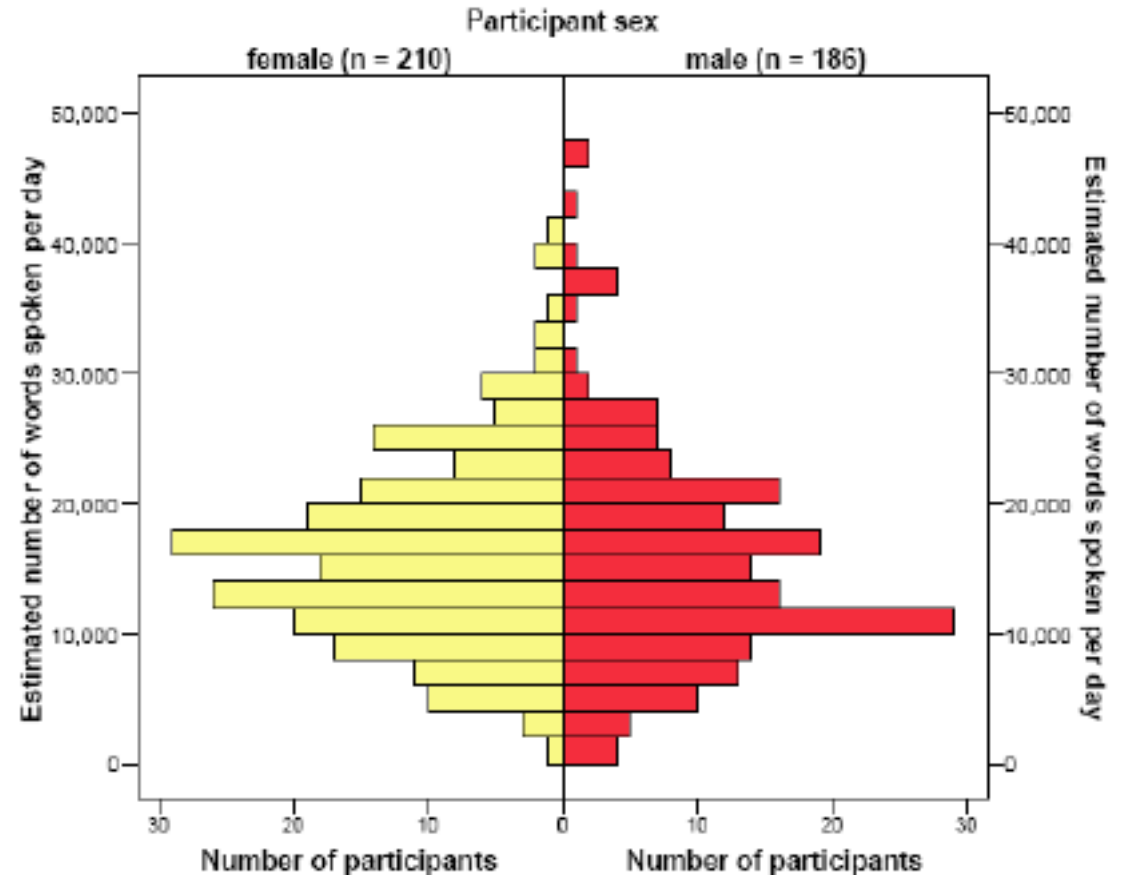
BREVIA

Are Women Really More Talkative Than Men?

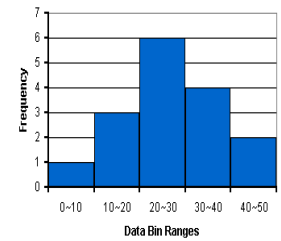
Histogram: estimated number of words spoken per day for female and male study participants (N=396)

Result: women and men both spoke about 16,000 words per day.

Careful with the tails ...
... always !

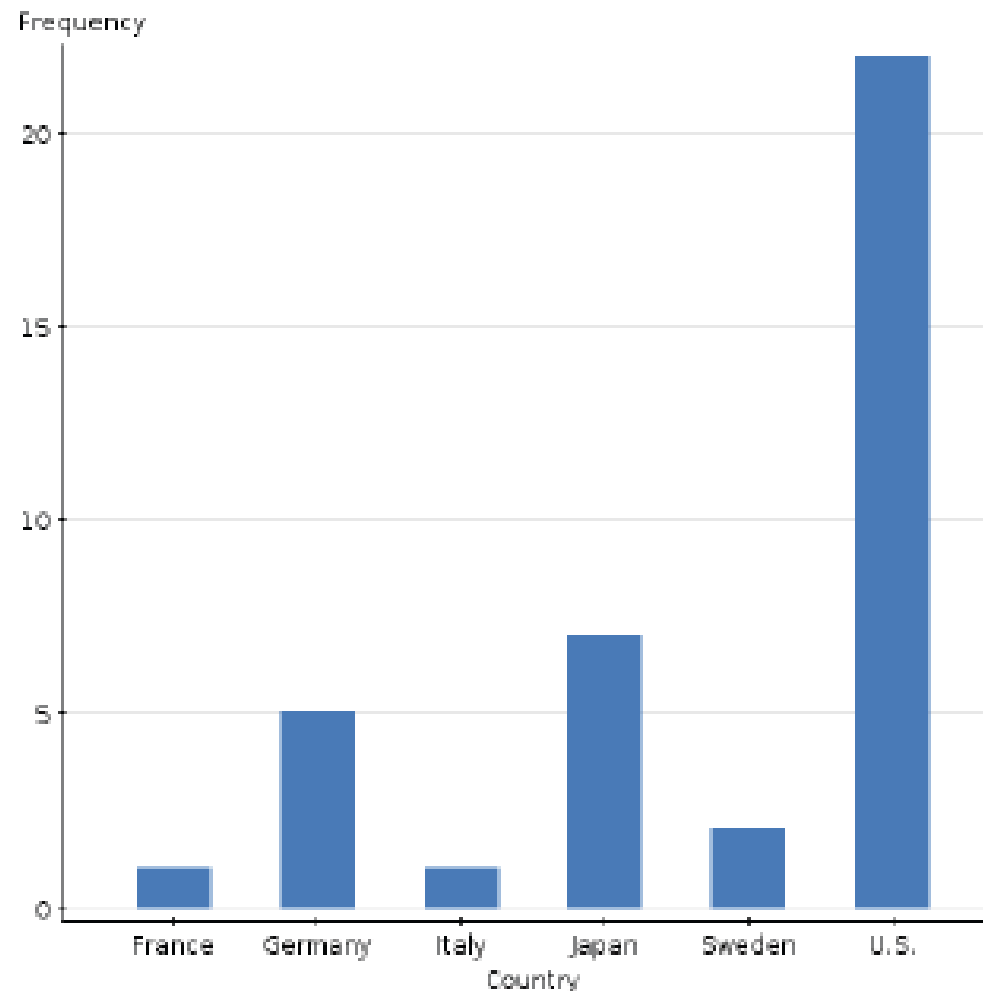


Non-numerical variables

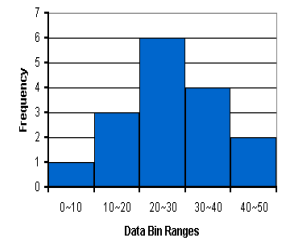


Variables of a distribution can also be not numerical, but any other quantity:

Country of origin of cars
in one sample:



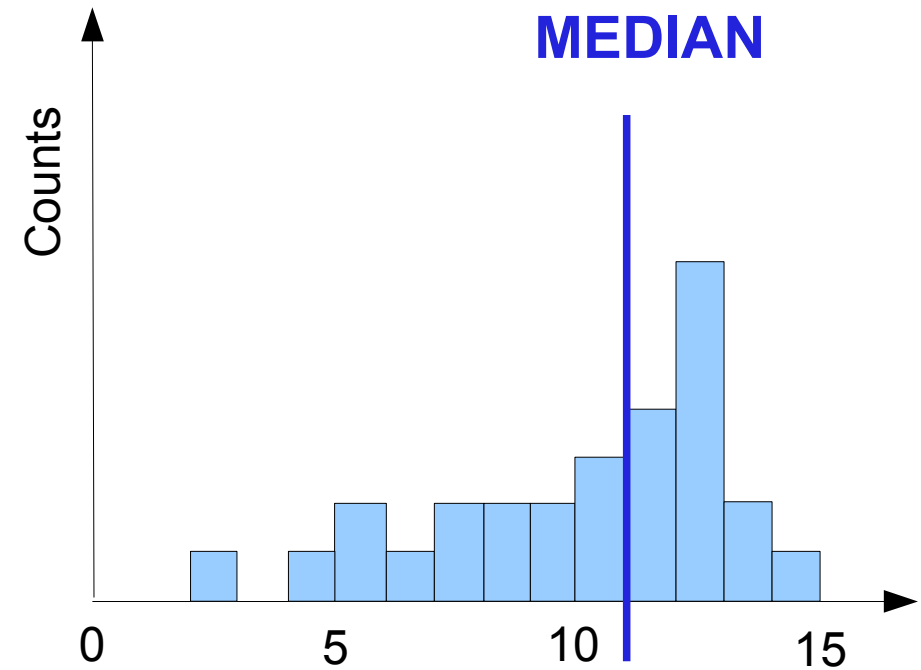
Estimators of a distribution



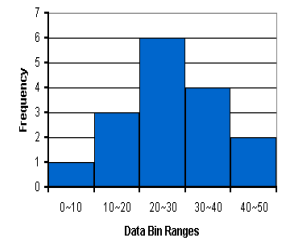
Estimate the center of a distribution:

MEDIAN:

Value dividing a sample into two sets, such that half of the data have a larger value



Estimators of a distribution

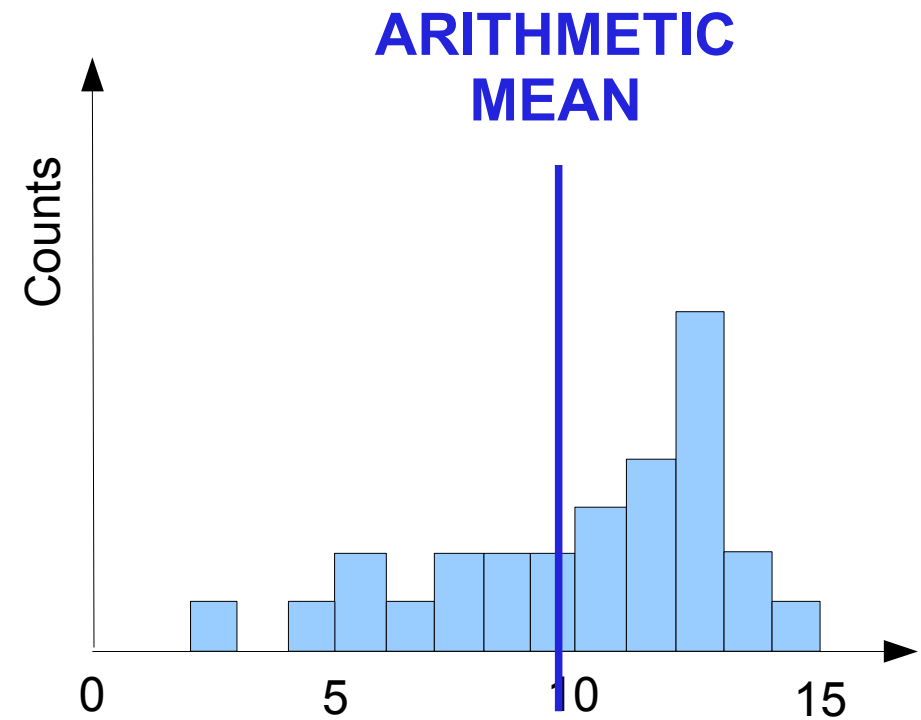


Estimate the center of a distribution:

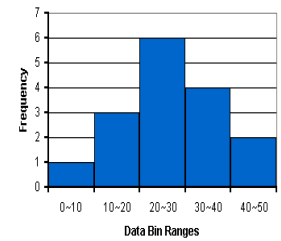
ARITHMETIC MEAN:

Sum of all observations of a sample divided by the number of observations

$$m_x = \frac{1}{N} \sum x_i$$



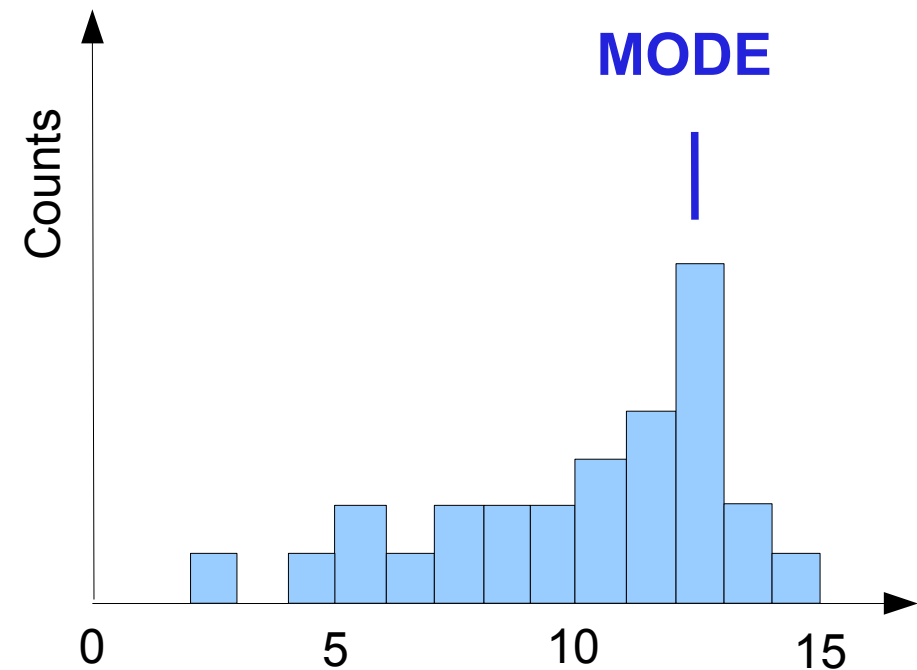
Estimators of a distribution



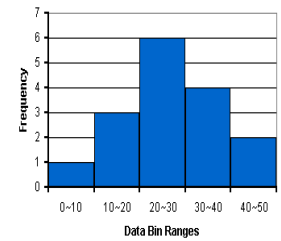
MODE (or modus):

The most probable value
(highest bin in distribution)

The definition is not really
unique (unimodal, bimodal
distributions)

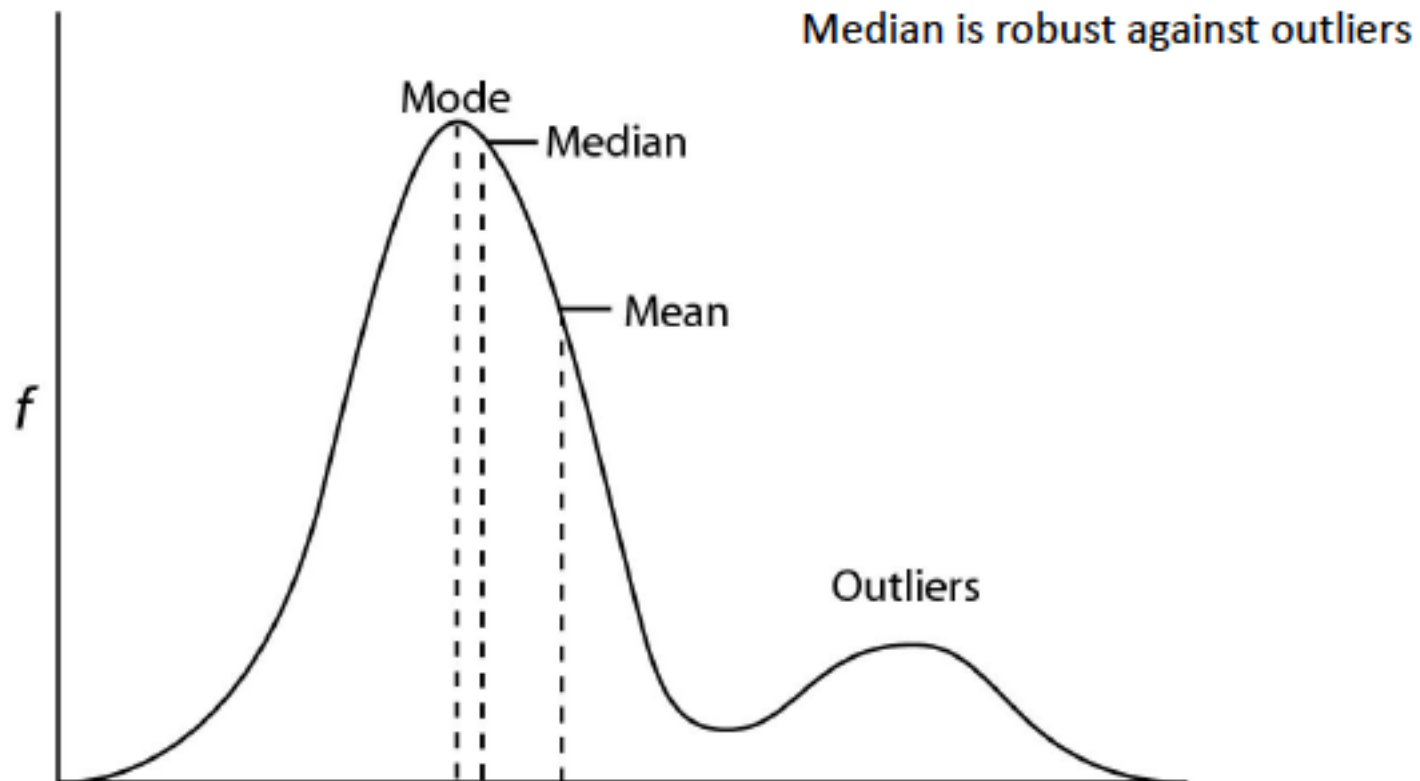


Estimators of a distribution

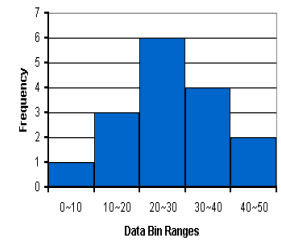


MEDIAN, MEAN, MODE:

Bimodal distribution



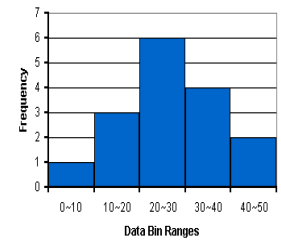
Simple exercises



Find the mean, median and mode of the following sets of numbers:

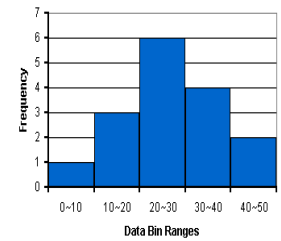
- A: 13, 18, 13, 14, 13, 16, 14, 21, 13
- B: -5, 3, -1, 3, 1, -1, 3, -2
- C: -5, 3, -1, 21, 1, -1, 3, -2
- D: 1, 2, 4, 7

Solution



- A:
 - Mean: 15
 - Median: 14
 - Mode: 13
- B:
 - Mean: 0.125
 - Median: 0
 - Mode: 3
- C:
 - Mean: 2.375
 - Median: 0
 - Mode: -1, 3
- D:
 - Mean: 3.5
 - Median: 3
 - Mode: none

More about means



We already mentioned the **arithmetic mean**

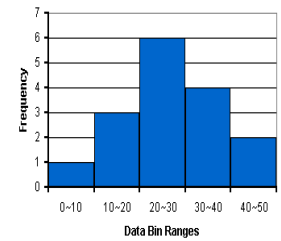
Definition:

$$\bar{x} = \frac{1}{N} \sum_i x_i$$

Examples:

- Average number of children in Germany is 2.3
- Average life expectation for men is 74, for women 78
- Average amount of semesters for physics studies in Heidelberg is 11.2

Weighted mean



Weighted mean

Definition:

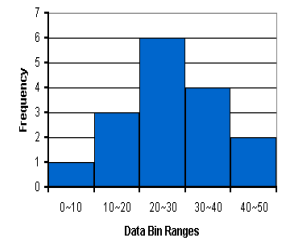
$$\bar{x} = \frac{1}{\sum w_i} \sum_i w_i x_i$$

Example:

- 5 measurements $\{x_1, x_2, x_3, x_4, x_5\}$ with different uncertainties $\{\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5\}$:

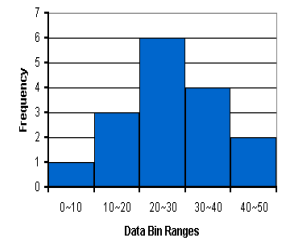
$$x = \frac{1}{\sum \frac{1}{\sigma_i^2}} \sum \frac{1}{\sigma_i^2} x_i$$

- The arithmetic mean is a special case of weighted mean ($w=1$)



- G.Cowan, “Statistical data analysis”, Clarendon Press, Oxford, 1998
Look also at: http://www.pp.rhul.ac.uk/~cowan/stat_course.html
- R.J.Barlow, “A Guide to the Use of Statistical Methods in the Physical Sciences”, John Wiley, 1989
- P.R.Bevington and D.K.Robinson, “Data reduction and error analysis for the physical sciences”, WBC/McGrow-Hill, 1992
- Previous edition of this course (source of much material!! Thanks Prof. Stephanie Hansmann-Menzemer!!)
<http://www.physi.uni-heidelberg.de/~menzemer/statistik10.html>

Next lecture



- Further characterization of distributions (width, standard deviation, variance, skewness, ...)
- Definition / interpretation of probability
 - Kolgomorov Axioms
- Random variables and probability densities
- Important distributions