Applied Mechanism Design and Big Data

→ sub-topic: Statistical Data Analysis

Max Baak

(with many thanks to Wouter Verkerke)

About me – Max Baak

- Job title: “Chief data scientist”
- @ KPMG (Big Data & Analytics team), since March 2015.

- Background in particle physics (2001-2015)
  - Interest in both theory and experiment
  - Past 7 years lived and worked at CERN (Geneva, Switzerland).
    - E.g. involved in/after Higgs boson discovery.
  - Graduated (PhD) at Nikhef.
  - During PhD, worked at Stanford for 3 years.

- Expert in statistical data analysis.

- Questions about the course?
  - Email: Baak.Max@kpmg.nl

Literature used for this course

Courses on Statistics

- *Practical statistics (for particle physicists)*
  https://indico.cern.ch/event/287744/timetable/#20140621
  Wouter Verkerke
- *Statistics*
  https://indico.cern.ch/event/243641/
  Kyle Cranmer

Books on probability theory

- *Statistical Data Analysis (Oxford Publishing)*
  Glen Cowan
- *Statistical methods in data analysis*
  W.J. Metzger
- *Superforcasting – The art and science of prediction (popular science)*
  P. Tetlock, D. Gardner

Wikipedia (of course!)

Credits & acknowledgements!

- Big thanks to **Wouter Verkerke** for recycling some of his statistics course material!

Wouter Verkerke

- Senior staff scientist at Nikhef.
  - Nikhef = Dutch institute for particle physics.
- Author of RooFit – toolkit for statistical data modeling!
  - Great software package data modeling!
  - See: http://root.cern.ch

Roadmap for this course

1. Statistics basics
   - Probability theory
   - Probability distributions

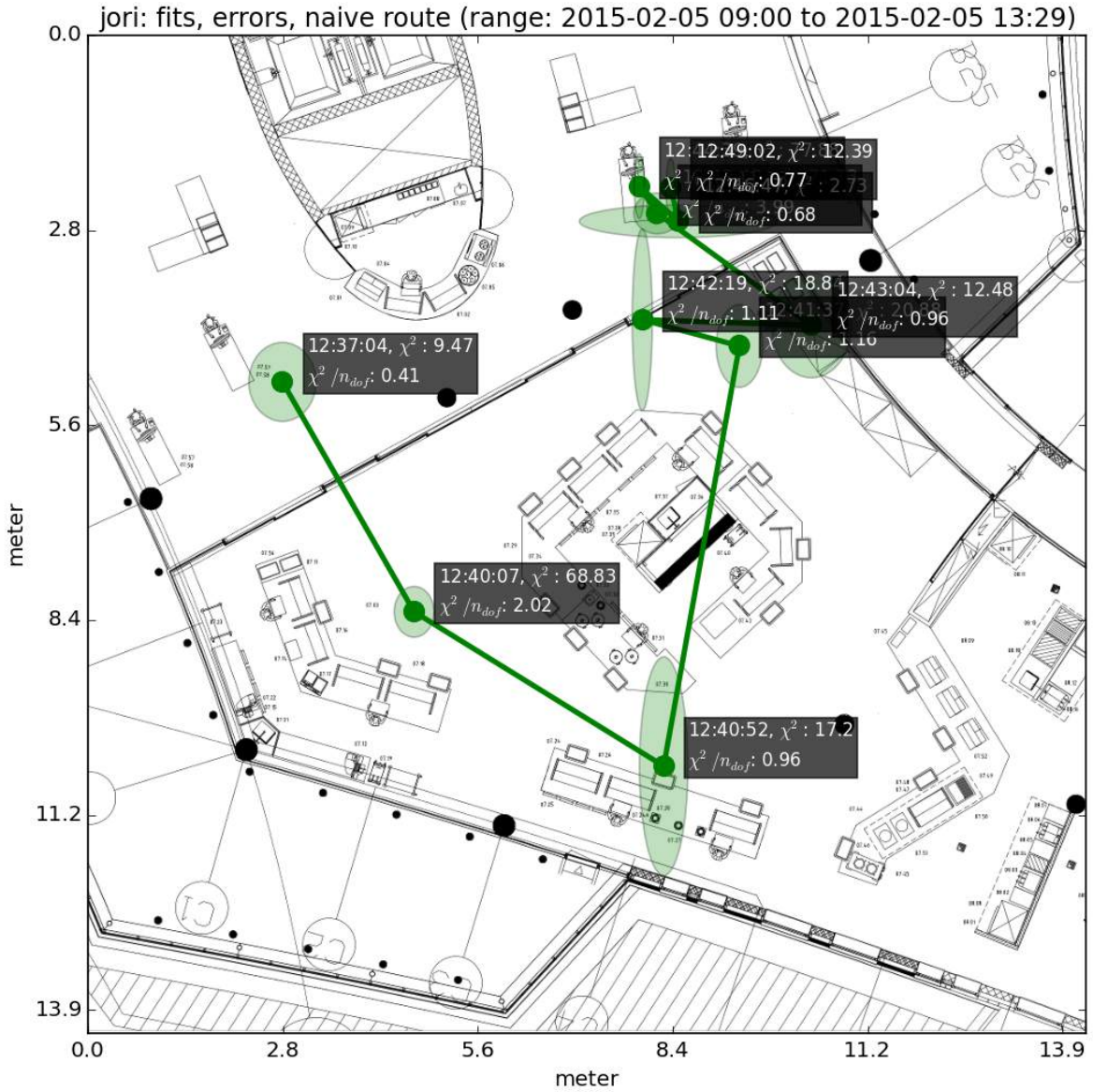
2. Parameter estimation

3. Pitfalls in (big) data analysis
   - Spurious correlations
   - Data quality assessment

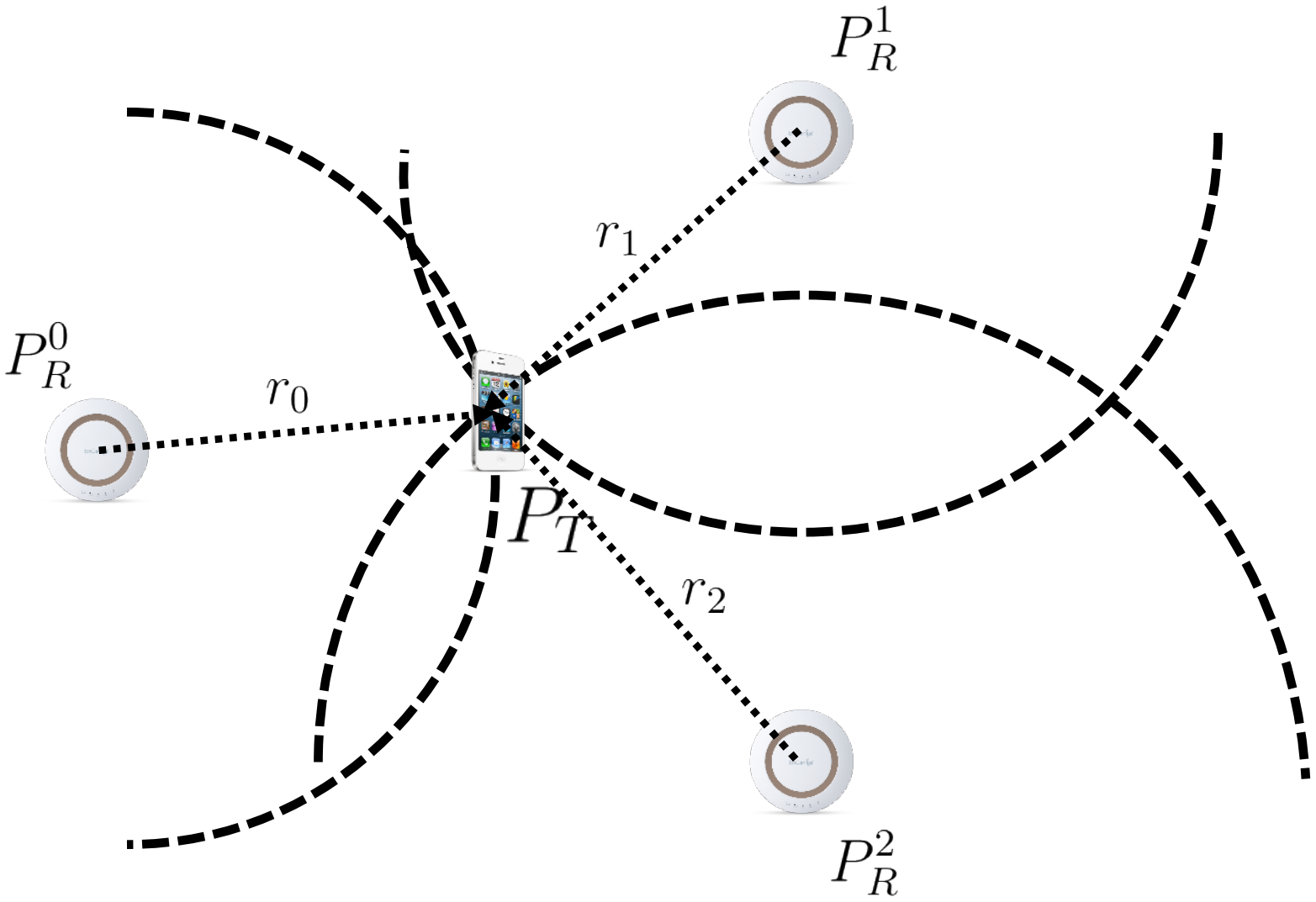
4. Hypothesis Testing

Introduction

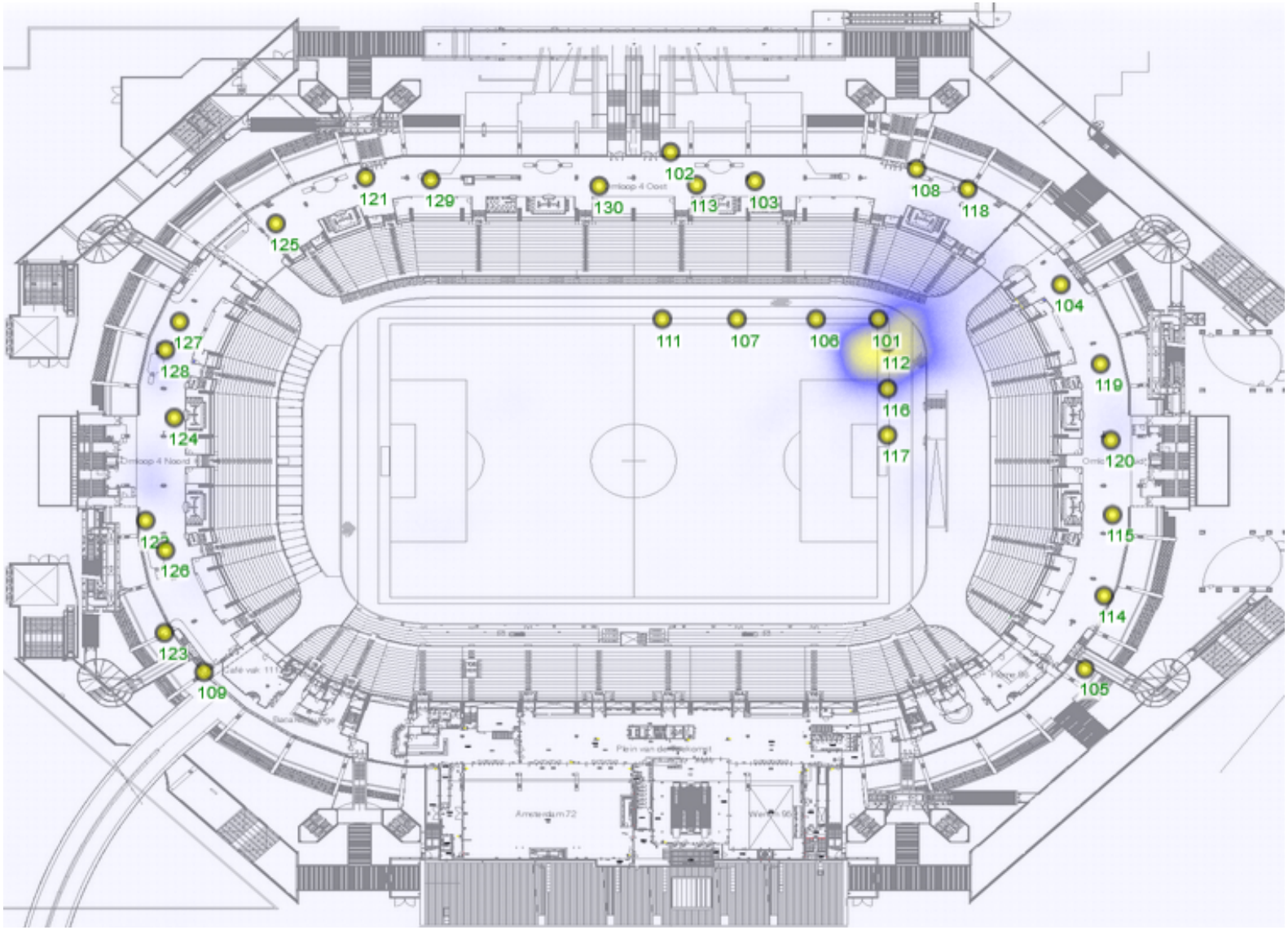
Tracking a smart phone through a restaurant



How does this work?



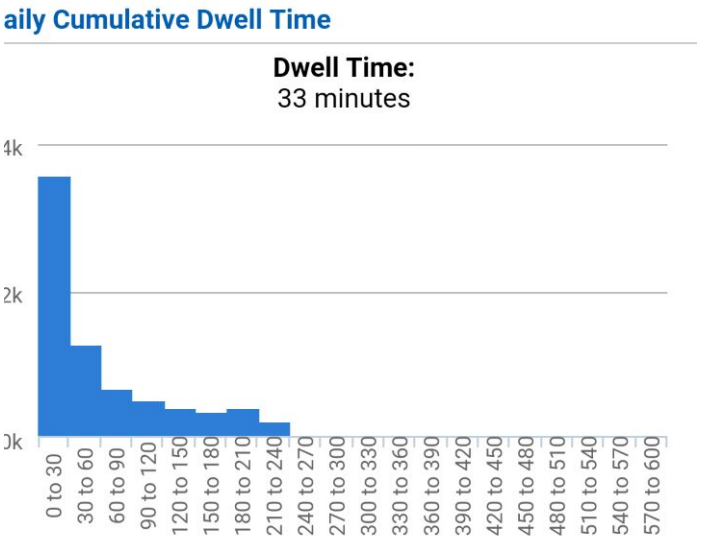
Example: Amsterdam Arena

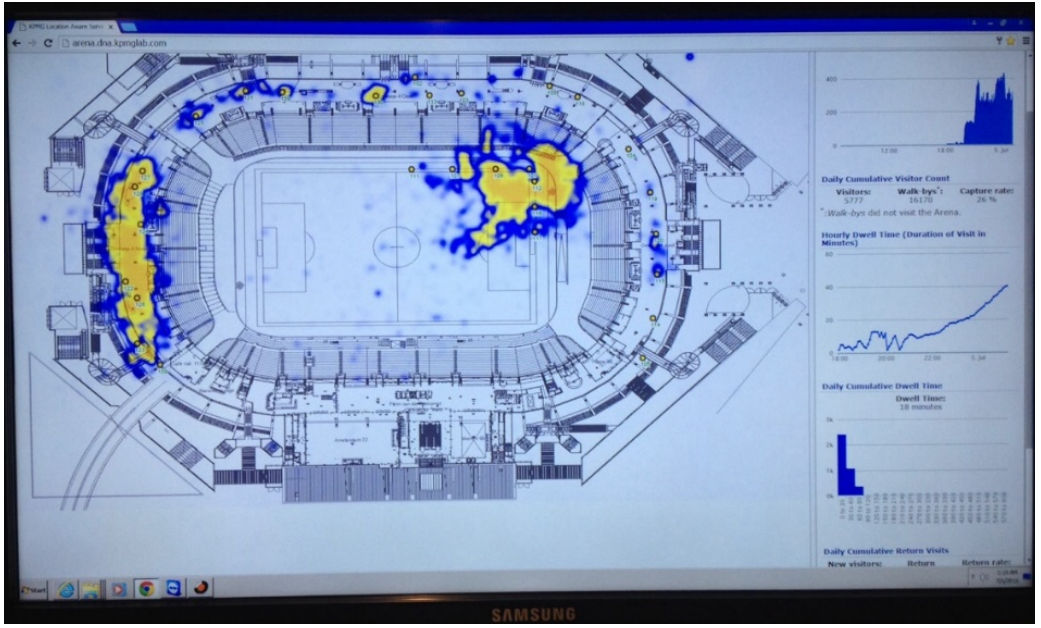


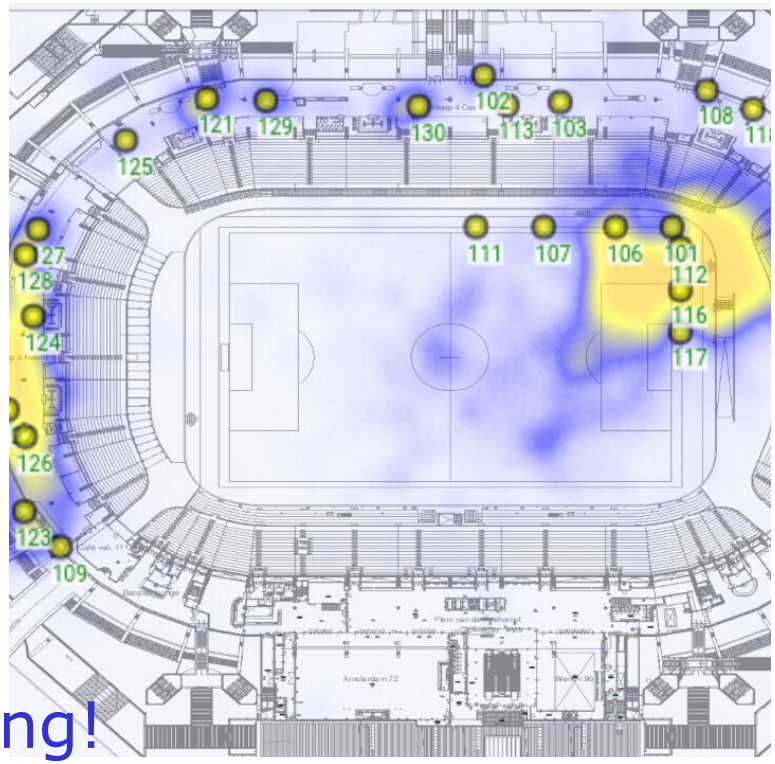
KPMG
cutting through complexity

Sensation White – 4 july 2015 (Amsterdam Arena)

Party all night …

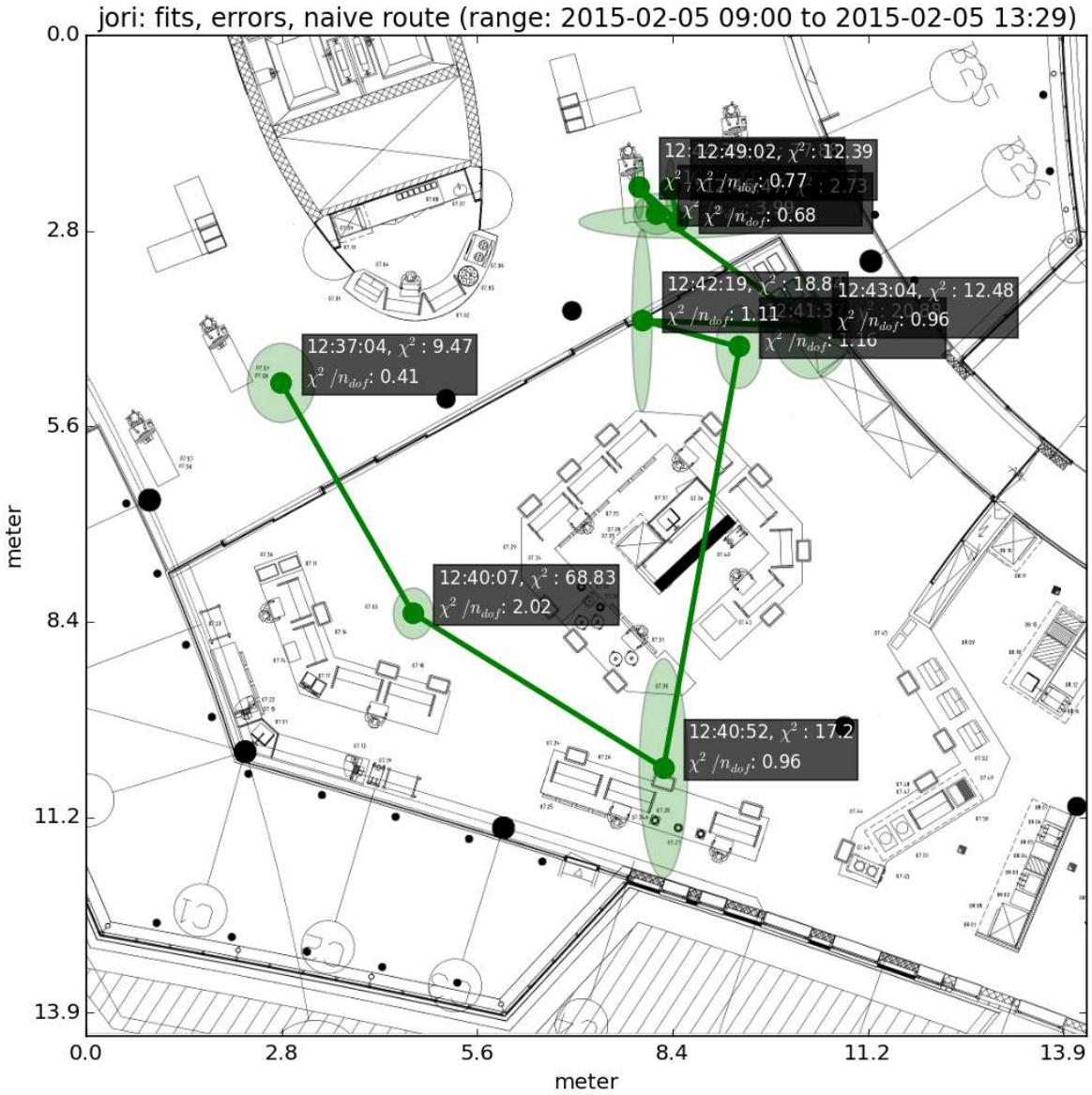






@ Amsterdam Arena: 600+ sensors available for wifi tracking!

Tracking a smart phone through a restaurant

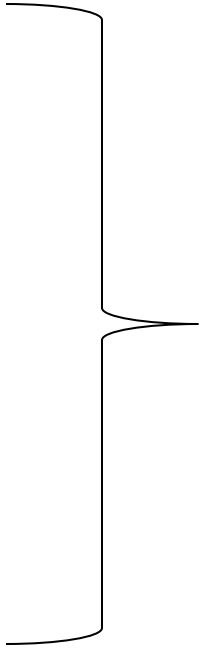


→ Part of Computer Practicum exercises!

Today's course

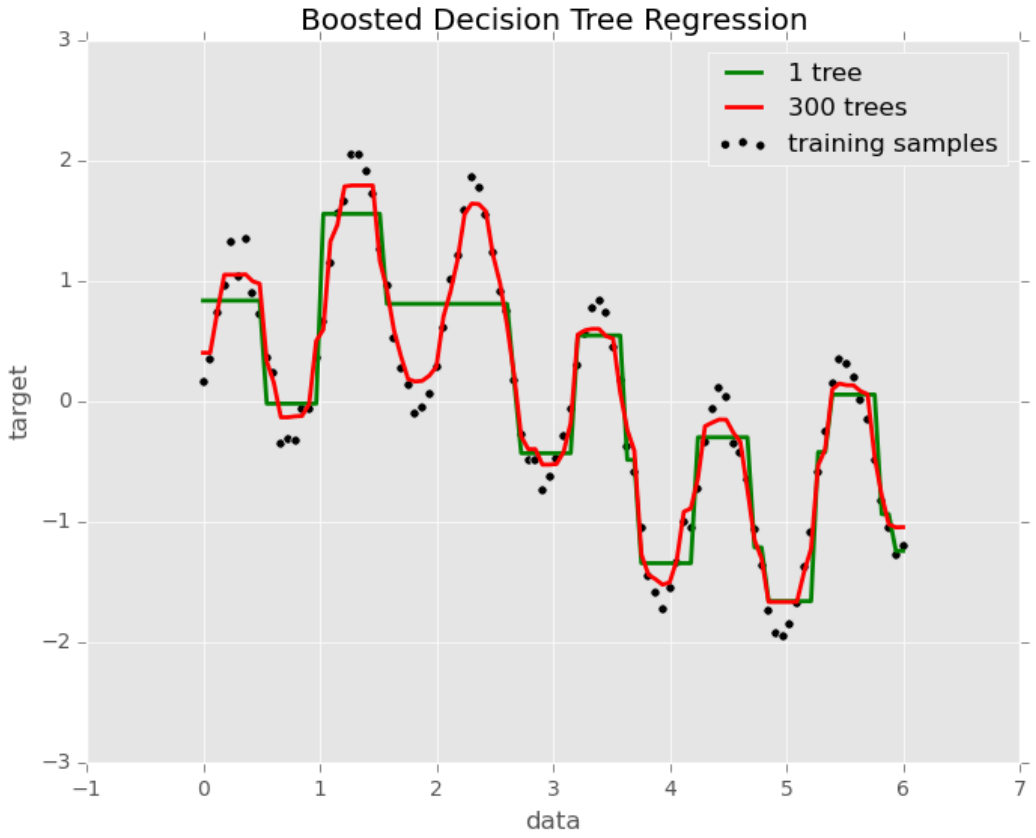
- Why data modelling "by hand" ? Advantages thereof.

- Introduction to basic statistics
- Probability distributions
- Measurement uncertainties
- Central limit theorem



Theory behind
and practical aspects
of data modeling

Simple data modelling example



- Building your own model to describe the data is benificial for many problems where “model is known”.
- Machine learning (ML) is not always the best solution ...
- *Goal of course: building and fitting your own model, and understanding the output.*

Similarities between ML and parametric data modeling

- Example of fitting decision tree to data.

- *Many practical similarities between finding optimal ML alg settings and fitting a parametric model to the data.*

```
# Create a random dataset
rng = np.random.RandomState(1)
X = np.sort(5 * rng.rand(80, 1), axis=0)
y = np.sin(X).ravel()
y[::5] += 3 * (0.5 - rng.rand(16))

# Fit regression model
clf_1 = DecisionTreeRegressor(max_depth=2)
clf_2 = DecisionTreeRegressor(max_depth=5)
clf_1.fit(X, y)
clf_2.fit(X, y)
```

1. model.fitTo(input_data)
2. Optimization = *(often) maximization of some (internal) function to find best parameter settings of ML alg to describe that data.*

http://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html#example-tree-plot-tree-regression-py

Differences b/n ML and data modeling by hand

- ML: think here of regression or event classification.

- Many problems are much more difficult than preceding example
  - Many observables, overwhelming background.
  - Various approaches possible …

- ML algorithm
  - Compactify information in multiple observables into one observable (e.g. output of neural network) → 'test statistic t(x)'
- Building your own parametric model: "Big fit"
  - Make explicit models of signal and/or background hypothesis in all observables and fit all data to estimate model parameters

Differences b/n ML and data modeling by hand

- Statistical sensitivity
  - ML usually throws away some information (but maybe very little.)
  - Big fit keeps all information. → highest precision.
- Feasibility
  - ML. Recent developments in machine learning make this a lot easier.
  - Big fit clearly most ambitious.
- What is the best you can do? Depends on the situation:
  - ML. Great for combining many input observables (>4).
  - Big fit.
    - The best approach when underlying model is known, and no "guessing" is required. E.g. model dictated by laws.
    - Very difficult to deal with many input observables (>4).

Advantages of modeling by hand (e.g. vs ML)

- Full control (over contents of the model).
  - You determine all ingredients of the model yourself.
    - E.g. What you want when underlying model is known.
- Assess the uncertainty on the fitted model
  - Can do error propagation of uncertainties on fit parameters.
- Simulation
  - Easy to test impact of parameters in model.
  - Often need for knobs that can be tuned.
    - E.g. test impact of policy changes on a population.
- Hypothesis testing
  - Likelihood fits to data have high(-est) sensitivity for classification.
  - Well-defined formulas to calculate p-values.

Basic Probability Theory

Video on probability of black holes

- See:
- http://www.cc.com/video-clips/hzqmb9/the-daily-show-with-jon-stewart-large-hadron-collider

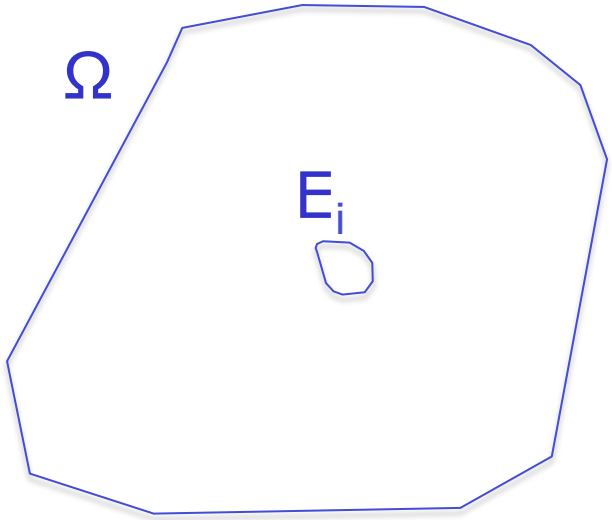
Probability

- The concept of probability is abstract to many people!
  - "30% chance of rain tomorrow in Amsterdam."
- Two primary states of mind:
  - A) event happens, or
  - B) event does not happen.
- Concept of "maybe" is an *acquired* state of mind.
  - People generally not concerned with percentages.
  - Example: the lottery. Either you win it, or you don't.
    - "The odds are fifty-fifty".
- "Maybe" very impractical from evolutionary perspective.
  - Often forced to make a choice between A) or B).

Probability

- Ω: set of all possible outcomes
  - Subspace E_i
- Frequentist probability:

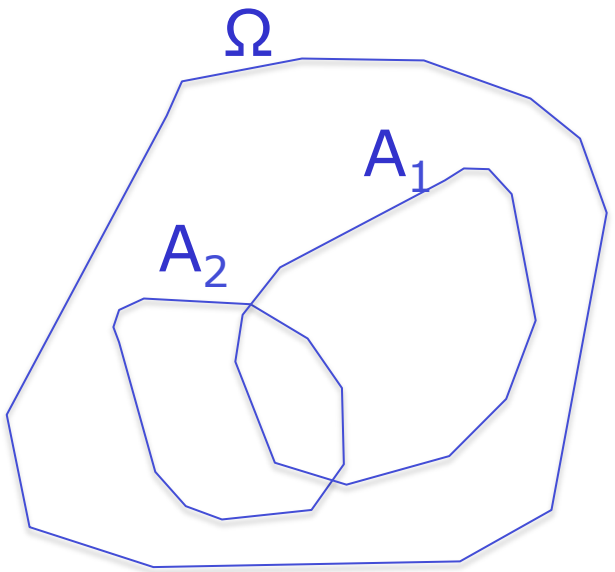
$$P(E_i) = \lim_{N \to \infty} \left(\frac{N_{Ei}}{N} \right)$$



- P(Ω) = 1
- 0 ≤ P(E) ≤ 1, E_i in Ω

Conditionality

Conditional probability:
P(E|A) = the probability of E, given the constraint A



- Renormalization: P(A|A)=1
- $P\left(A_2 \mid A_1\right) = P\left(A_2 \cap A_1 \mid A_1\right)$
- Ratio is constant: $\frac{P\left(A_2 \cap A_1 \mid A_1\right)}{P\left(A_1 \mid A_1\right)} = \frac{P\left(A_2 \cap A_1\right)}{P\left(A_1\right)}$
- It follows that: $P\left(A_2 \mid A_1\right) \times P\left(A_1\right) = P\left(A_2 \cap A_1\right)$

Bayes' theorem

- $A_1 \cap A_2 = A_2 \cap A_1 \quad \Rightarrow \quad P(A_1 \cap A_2) = P(A_2 \cap A_1)$
- Using: $P(A_2 \mid A_1) \times P(A_1) = P(A_2 \cap A_1)$
- … results in: $P(A_2 \mid A_1) P(A_1) = P(A_1 \mid A_2) P(A_2)$

put differently: $$P(A_2 \mid A_1) = \frac{P(A_1 \mid A_2)}{P(A_1)} P(A_2)$$

Bayes' theorem - Example

$$P(T \mid pos) = \frac{P(pos \mid T)}{P(pos)} P(T)$$

- Given a city of 1 million inhabitants,
  let there be 100 terrorists (and 999,900 non-terrorists)
  → P(T) = 0.0001
- A terrorist test has 99% accuracy
  It will correctly identify a terrorist 99% of the time, and correctly give a negative result 99% of the time.

- Q: If you get a positive result, what are the odds that you actually found a terrorist, P(T|pos) ?
- *P(pos) = 0.0001*0.99 + 0.9999*0.01 = 0.0101*
- P(T|pos) = 0.0098
  - I.e. less than 1%. Entirely dominated by false-positive rate.
  - Mind you, this assumes people are picked out at random!

Frequency interpretation

Frequentist interpretation of probability:

$$P(E_i) = \lim_{N \to \infty} \left(\frac{N_i}{N} \right)$$

a.k.a. Emperical probability

1. The experiment must be repeatable, under identical conditions!
   - "What's the probability that it will rain tomorrow?" Meaningless question to Frequentists!
2. $P(E_i)$ depends on the "ensemble", i.e. N repetitions of the experiment

Bayesian interpretation

- a.k.a. subjective probability
- Based on Bayes' theorem: $P(A \mid B)P(B) = P(B \mid A)P(A)$
- Interpretation:
  - A = theory or interpretation
  - B = result or observation
- Then:

$$P(theory \mid result) = \frac{P(result \mid theory)}{P(result)} P(theory)$$

"Belief" in theory *after* result

Measurement

"Belief" in theory *before* result

Bayes' postulate

- Want to measure parameter of nature, λ
- … by doing an experiment with outcome Z

$$P_{posterior}\left(\lambda \mid Z\right) = \frac{P\left(Z \mid \lambda\right)}{P\left(Z\right)} P_{prior}\left(\lambda\right)$$

- *Bayes' postulate:*
  If ignorant about $P_{prior}\left(\lambda\right)$,
  take all values of λ to be equally probable

Bayes' postulate

Issues:

1. Choice of flat prior is truly guess
2. A different choice of prior gives a different posterior!

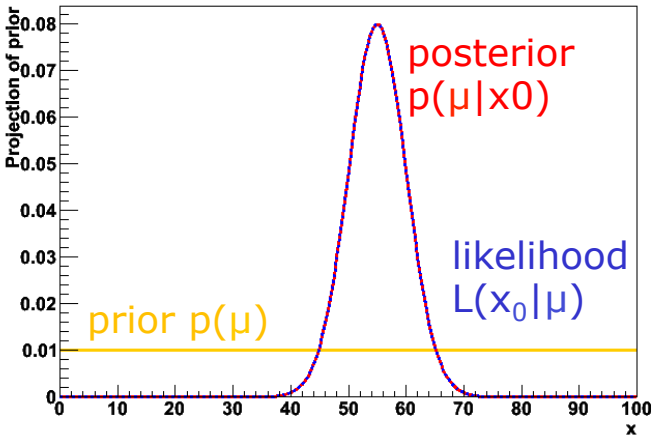
Typical objection:

- $P_{posterior}$ usually converges to same value/distribution, irrespective of chosen P_{prior}, if measurement(s) are strong

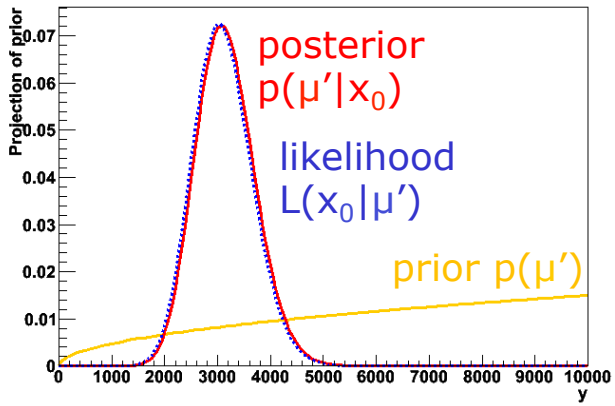
Choosing Priors

- Example of uniform vs quadratic prior:

distribution in **μ**



distribution in **μ²**



- Right-hand side:
  $P_{posterior}$ is the almost the same b/c the measurement dominates.

Frequentist or Bayesian?

- Suppose you have measured the electron mass to be: 520 ± 10 keV/c^2
  - "The mass of the electron is between 510 and 530 keV/c^2 with 68% probability."
    - **(Bayesian statement)**

- Frequentist: electron has definite mass, we just do not know the exact value
  - "The mass of the electron is between 510 and 530 keV/c^2 with 68% *confidence*."

- In interpreting experimental results, we are often Bayesians

$$P\left(m_{obs} \mid m_{true}\right) \leftrightarrow P\left(m_{true} \mid m_{obs}\right)$$

Frequentist or Bayesian?

- "A Frequentist is a person whose long-run ambition is to be wrong 5% of the time."
- "A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule."
- (Sources unknown)
- *→ It is important to be able to work with either definition of probability, and to know which one you are using!*

Basic Statistics

Describing your data – the Average

- Given a set of *unbinned* data (measurements)

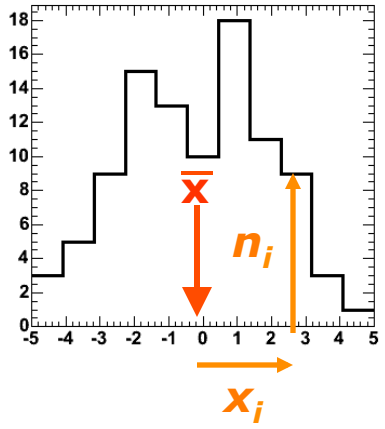
  $\{ x_1, x_2, \ldots, x_N \}$

  then the mean value of x is

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- For *binned* data

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} n_i x_i$$



  - where n_i is bin count and x_i is bin center
  - Unbinned average more accurate due to rounding

Describing your data – Variance

- *Variance V(x)* of x expresses how much ***x*** is liable to vary from its mean value $\bar{x}$

$$
\begin{aligned}
V(x) &= \frac{1}{N}\sum_i (x_i - \bar{x})^2 \\
&= \frac{1}{N}\sum_i (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \frac{1}{N}\sum_i x_i^2 - \frac{1}{N}2\bar{x}\sum_i x_i + \frac{1}{N}\bar{x}^2\sum_i 1) \\
&= \overline{x^2} - 2\bar{x}^2 + \bar{x}^2 \\
&= \overline{x^2} - \bar{x}^2
\end{aligned}
$$

- ***Standard deviation, or root-mean-squared (RMS):***

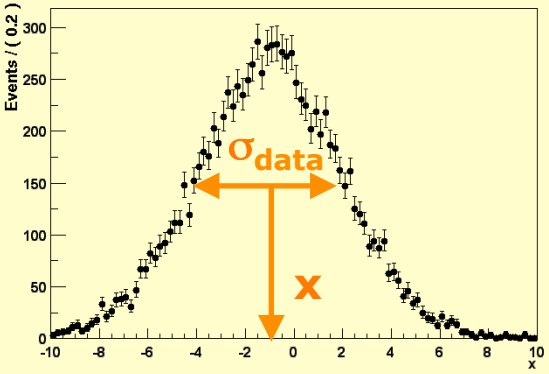
$$\sigma \equiv \sqrt{V(x)} = \sqrt{\overline{x^2} - \bar{x}^2}$$

Different definitions of the Standard Deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_i (x^2 - \bar{x})^2}$$ is the S.D. of the ***data sample***

- Presumably our data was taken from a parent distributions which has mean μ and S.D. σ

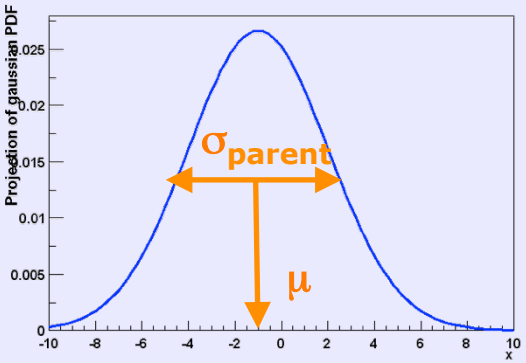
Data Sample



$\bar{x}$ – mean of our sample

σ – S.D. of our sample

Parent Distribution
(from which data sample was drawn)



μ – mean of our parent dist

σ – S.D. of our parent dist

Beware Notational Confusion!

More than one observable

- Given *2 observables x,y* and a dataset consisting of pairs of numbers

$$\{ (x_1,y_1),\ (x_2,y_2), \ldots (x_N,y_N) \}$$

- Definition of $\overline{x}$, $\overline{y}$, σ_x, σ_y as usual
- In addition, any *dependence between x,y* described by the ***co***variance

$$\begin{aligned} \mathrm{cov}(x,y) &= \frac{1}{N}\sum_i (x_i - \bar{x})(y_i - \bar{y}) \\ &= \overline{(x - \bar{x})(y - \bar{y})} \\ &= \overline{xy} - \bar{x}\,\bar{y} \end{aligned}$$

(has dimension D(x)D(y))

- The dimensionless ***correlation*** coefficient is defined as $\rho = \dfrac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} \in [-1,+1]$

Correlation & covariance in >2 variables

- Concept of covariance, correlation is easily extended to arbitrary number of variables

$$\mathrm{cov}(x_{(i)}, x_{(j)}) = \overline{x_{(i)} x_{(j)}} - \bar{x}_{(i)} \bar{x}_{(j)}$$

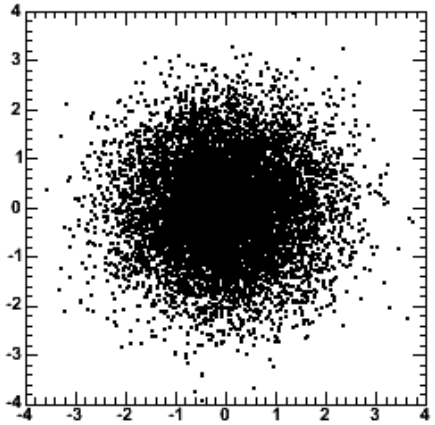
- so that $V_{ij} = \mathrm{cov}(x_{(i)}, x_{(j)})$ takes the form of a *n x n symmetric matrix*

- This is called the ***covariance matrix***, or ***error matrix***
- Similarly the correlation matrix becomes

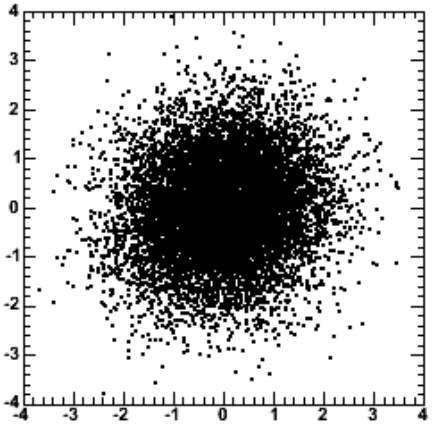
$$\rho_{ij} = \frac{\mathrm{cov}(x_{(i)}, x_{(j)})}{\sigma_{(i)} \sigma_{(j)}} \longrightarrow V_{ij} = \rho_{ij} \sigma_i \sigma_j$$

Visualization of correlation

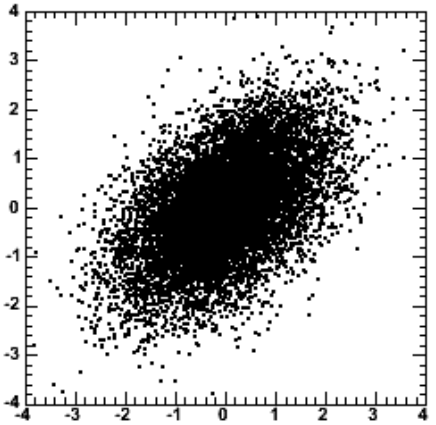
$\rho = 0$



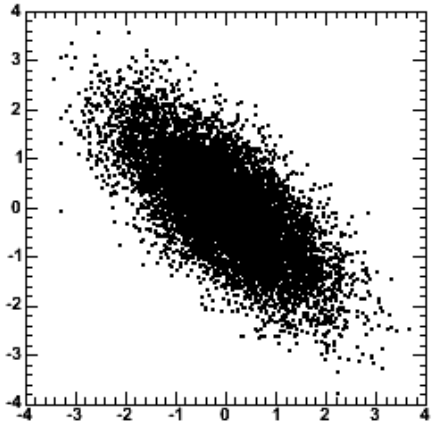
$\rho = 0.1$



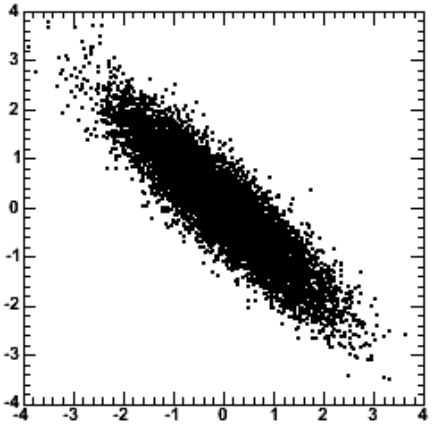
$\rho = 0.5$



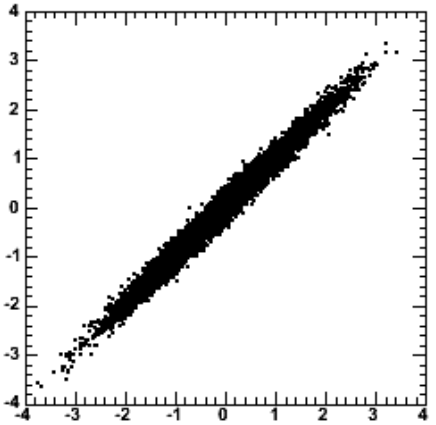
$\rho = -0.7$



$\rho = -0.9$



$\rho = 0.99$



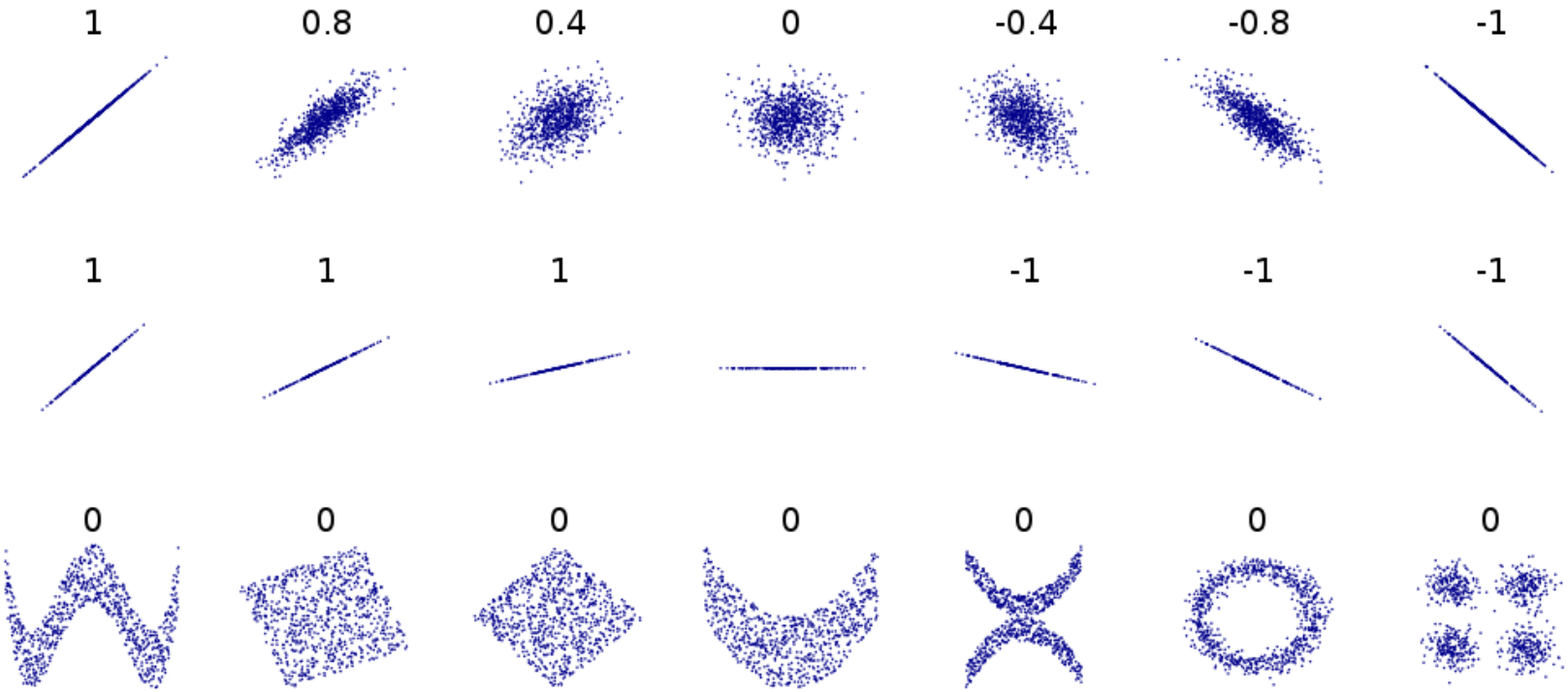
Guess the correlation!

- guessthecorrelation.com

- Homework exercise (optional):
submit your top score to Philip Rutten

Linear vs non-linear correlations

- Correlation coeffients used here are (linear) Pearson product-moment correlation coefficient
- Data can have more subtle (non-linear correlations) that expressed in these coefficient



- Always check correlation by eye!

Other measures of (non-linear) correlation

1. Correlation ratio η^2:

$$\eta^2(Y|X) = \frac{\sigma_{E(Y|X)}}{\sigma_Y},$$

$$E(Y|X) = \int y\ P(y|x)\ dy,$$

2. Mutual information I:

$$I(X,Y) = \sum_{X,Y} P(X,Y) \ln \frac{P(X,Y)}{P(X)P(Y)}$$

Related to entropy

- Examples from: http://tmva.sourceforge.net/docu/TMVAUsersGuide.pdf
- Functions available in TMVA library
  Or search on google ☺

Applied examples

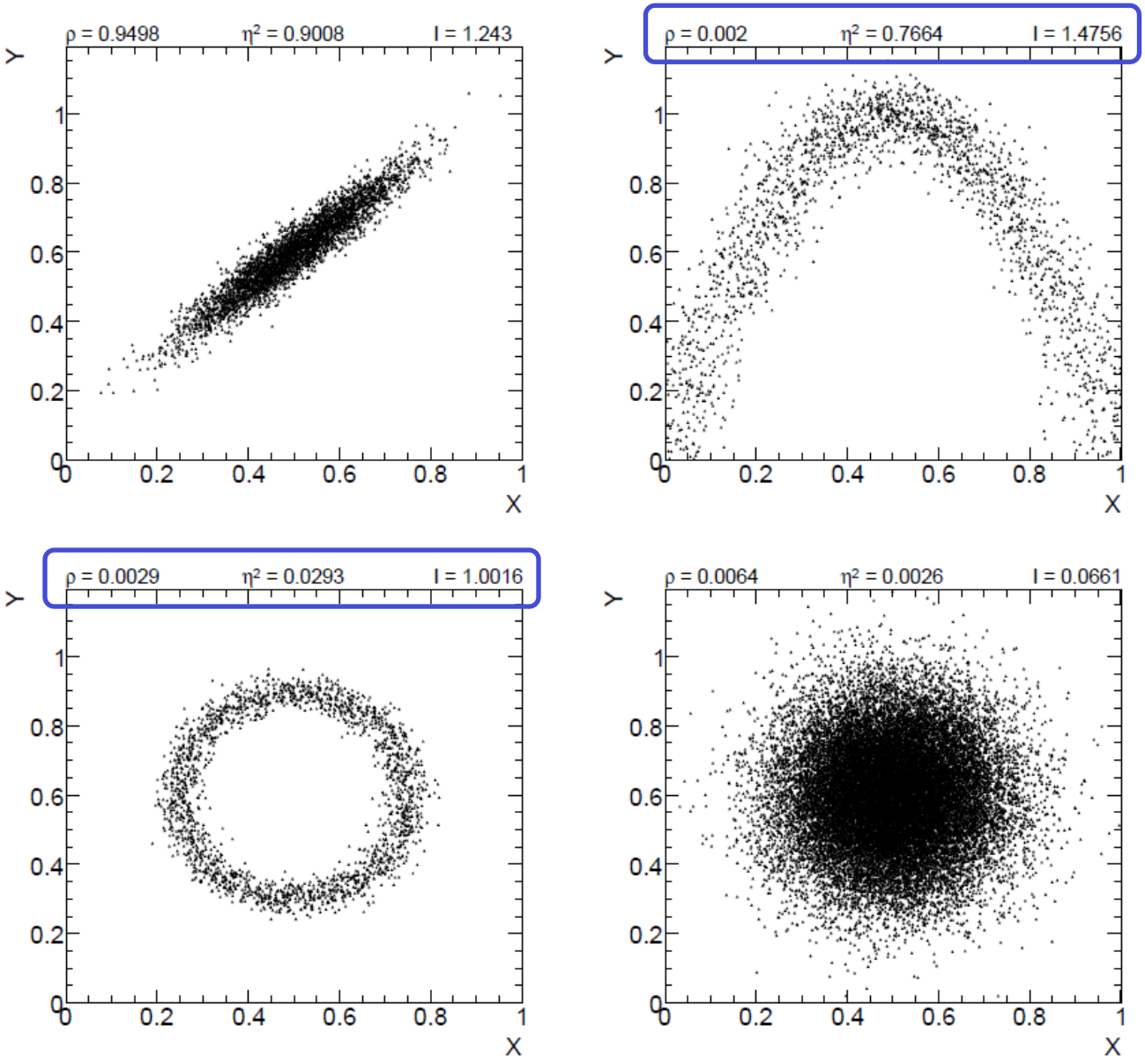


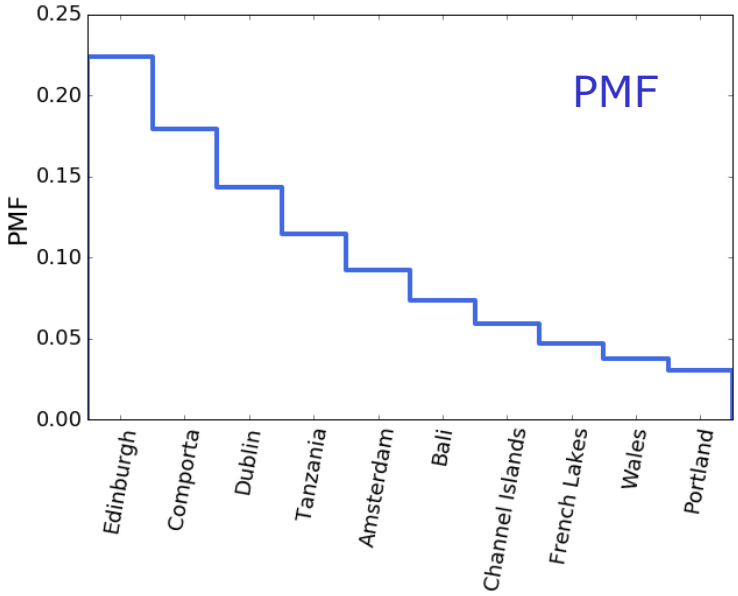
Figure 8: Various types of correlations between two random variables and their corresponding values for the correlation coefficient ρ, the correlation ratio η, and mutual information I. Linear relationship (upper left), functional relationship (upper right), non-functional relationship (lower left), and independent variables (lower right).

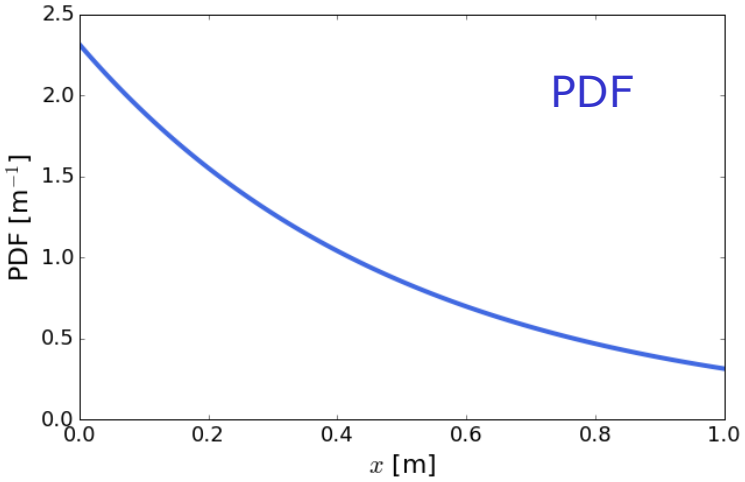
- Accidental correlations → Discussed in lecture 3.

Properties of distributions

Distribution of a random variable

- Statistical properties of random variables in dataset described by a "distribution function"
- Discrete variables: Probability Mass Function (PMF)
  - Integer numbers, strings, …
  - Probability to observe a particular value in a dataset entry
- Continuous variables: Probability Density Function (PDF)
  - Floating-point numbers
  - Probability per "unit of variable" to observe in a dataset entry
  - Probability to observe value in a given range: integral over range





Distribution of a random variable

- Expectation values of statistical properties determined by distribution:

$$\langle \bar{x} \rangle \equiv \int x \cdot P(x)\, \mathrm{d}x$$

$$\langle V(x) \rangle \equiv \int (x - \bar{x})^2 \cdot P(x)\, \mathrm{d}x$$

- Averages go to expectation values if "experiment" that produced the dataset is repeated many (n) times:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{e=0}^{n} \bar{x}_e = \langle \bar{x} \rangle$$

$$\langle \bar{x} \rangle \equiv \int x \cdot P(x)\, \mathrm{d}x$$

$$\langle V(x) \rangle \equiv \int (x - \bar{x})^2 \cdot P(x)\, \mathrm{d}x$$

- Extends to distributions of more than one variable

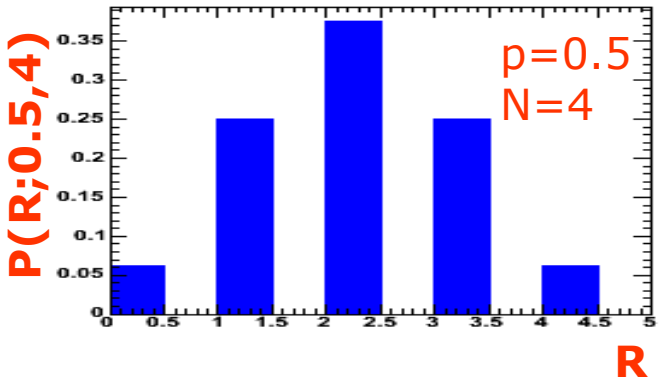
Basic Distributions – The binomial distribution

- Simple experiment – Drawing marbles from a bowl
  - Bowl with marbles, fraction **p** are black, others are white
  - Draw **N** marbles from bowl, *put marble back after each drawing*
  - Distribution of **R** black marbles in drawn sample:

Probability of a specific outcome e.g. 'BBBWBWW'

Number of equivalent permutations for that outcome

$$P(R;p,N) = p^R(1-p)^{N-R}\frac{N!}{R!(N-R)!}$$

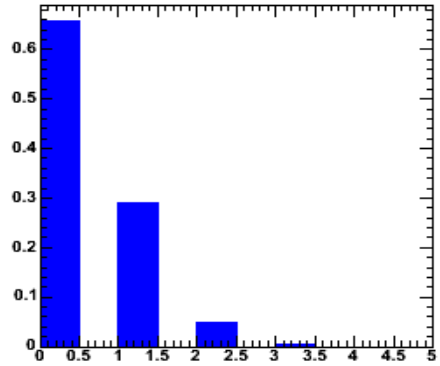


Binomial distribution

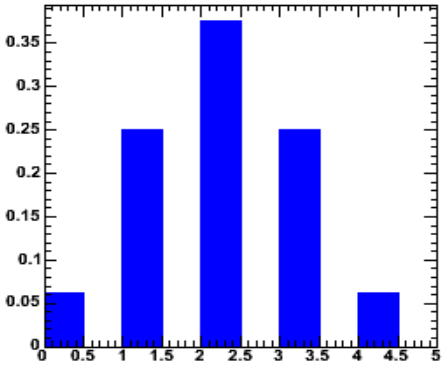
Properties of the binomial distribution

- Mean: $\langle r \rangle = n \cdot p$
- Variance: $V(r) = np(1-p) \quad \Rightarrow \quad \sigma = \sqrt{np(1-p)}$

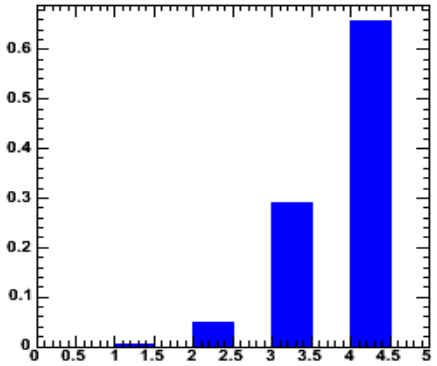
p=0.1, N=4



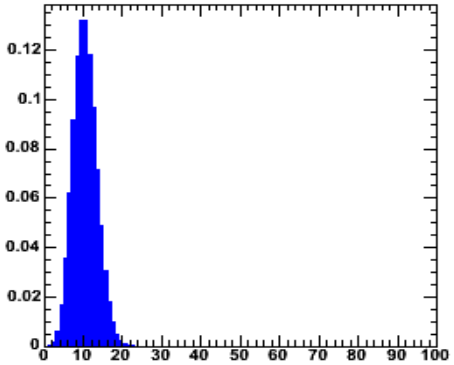
p=0.5, N=4



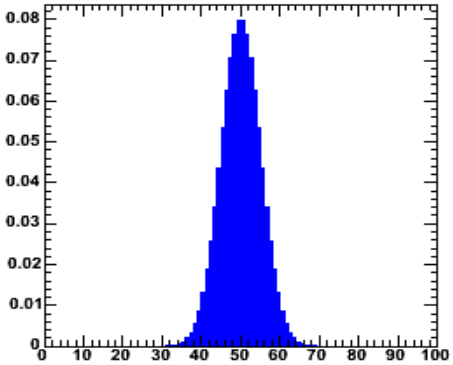
p=0.9, N=4



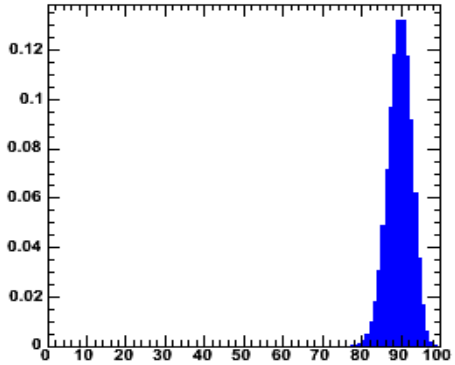
p=0.1, N=1000



p=0.5, N=1000



p=0.9, N=1000



Basic Distributions – the Poisson distribution

- Sometimes we don't know the equivalent of the number of drawings. Examples:
  - **Number of golden-colored cars passing by on high way.**
  - **Geiger counter to measure radioactive decay.**
  - Sharp events occurring in a (time) continuum

- What distribution to we expect in measurement over fixed amount of time?
  - Divide time interval λ in n finite chunks,
  - Take binomial formula with p=λ/n and let n→∞

$$P(r;\lambda/n,n) = \frac{\lambda^r}{n^r}\left(1-\frac{\lambda}{n}\right)^{n-r}\frac{n!}{r!(n-r)!}$$

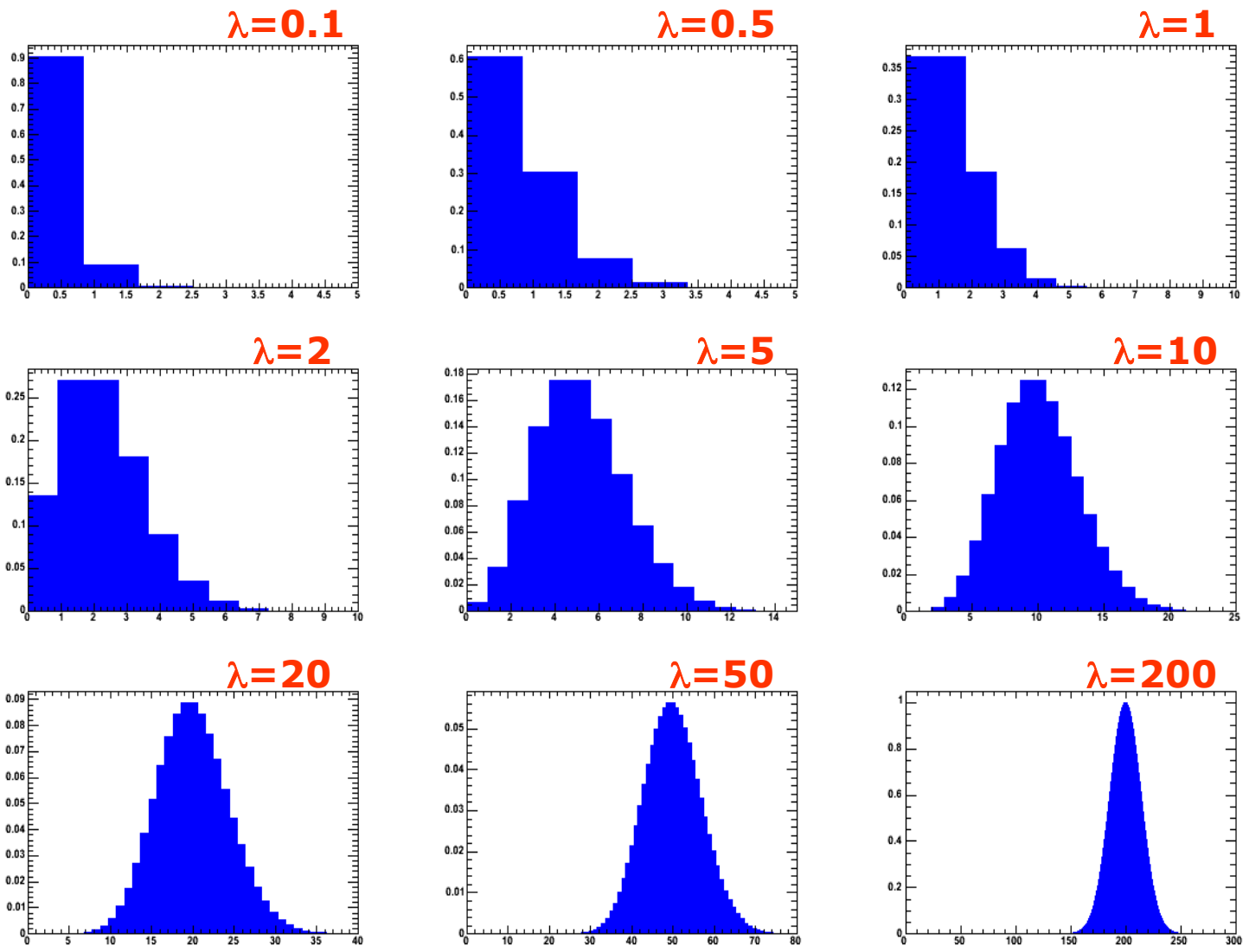
$$\lim_{n\to\infty}\frac{n!}{(n-r)!} = n^r,$$

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n-r} = e^{-\lambda}$$

$$P(r;\lambda) = \frac{e^{-\lambda}\lambda^r}{r!}$$

←**Poisson distribution**

Properties of the Poisson distribution



More properties of the Poisson distribution $P(r;\lambda) = \frac{e^{-\lambda}\lambda^r}{r!}$

- Mean, variance: $\langle r \rangle = \lambda$

$$V(r) = \lambda \quad \Rightarrow \quad \sigma = \sqrt{\lambda}$$

- Convolution of 2 Poisson distributions is also a Poisson distribution with $\lambda_{ab}=\lambda_a+\lambda_b$

$$P(r) = \sum_{r_A=0}^{r} P(r_A;\lambda_A)P(r-r_A;\lambda_B)$$

$$= e^{-\lambda_A}e^{-\lambda_B}\sum \frac{\lambda_A^{r_A}\lambda_B^{r-r_A}}{r_A!(r-r_A)!}$$

$$= e^{-(\lambda_A+\lambda_B)}\frac{(\lambda_A+\lambda_B)^r}{r!}\sum_{r_{A=0}}^{r}\frac{r!}{(r-r_A)!}\left(\frac{\lambda_A}{\lambda_A+\lambda_B}\right)^{r_A}\left(\frac{\lambda_B}{\lambda_A+\lambda_B}\right)^{r-r_A}$$

$$= e^{-(\lambda_A+\lambda_B)}\frac{(\lambda_A+\lambda_B)^r}{r!}\left(\frac{\lambda_A}{\lambda_A+\lambda_B}+\frac{\lambda_B}{\lambda_A+\lambda_B}\right)^{r}$$

$$= e^{-(\lambda_A+\lambda_B)}\frac{(\lambda_A+\lambda_B)^r}{r!}$$

Example of r = 0: Snowden formula

Volkskrant 27 jan '16

Basic Distributions – The Gaussian distribution

- Look at *Poisson distribution* in limit of *large N*

$$P(r;\lambda) = e^{-\lambda}\frac{\lambda^r}{r!}$$

Take log, substitute, r = λ + x, and use $\ln(r!) \approx r\ln r - r + \ln\sqrt{2\pi r}$

$$\ln(P(r;\lambda)) = -\lambda + r\ln\lambda - (r\ln r - r) - \ln\sqrt{2\pi r}$$

$$= -\lambda + r\left[\ln\lambda - \ln(\lambda(1+\frac{x}{\lambda}))\right] + (\lambda + x) - \ln\sqrt{2\pi\lambda}$$

$$\approx x - (\lambda + x)\left(\frac{x}{\lambda} - \frac{x^2}{2\lambda^2}\right) - \ln(2\pi\lambda)$$

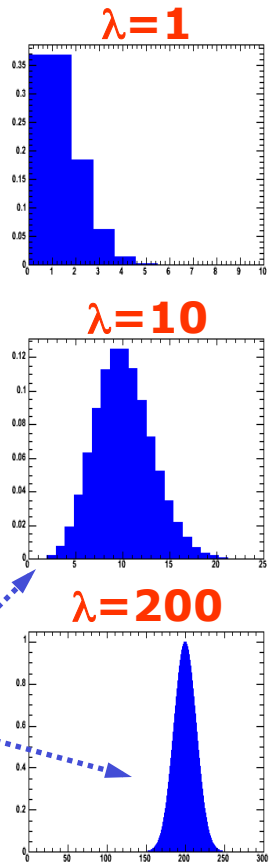
$\ln(1+z) \approx z - z^2/2$

$$\approx \frac{-x^2}{2\lambda} - \ln(2\pi\lambda)$$

Take exp

$$P(x) = \frac{e^{-x^2/2\lambda}}{\sqrt{2\pi\lambda}}$$

Familiar Gaussian distribution,
(approximation reasonable for N>10)



Properties of the Gaussian distribution

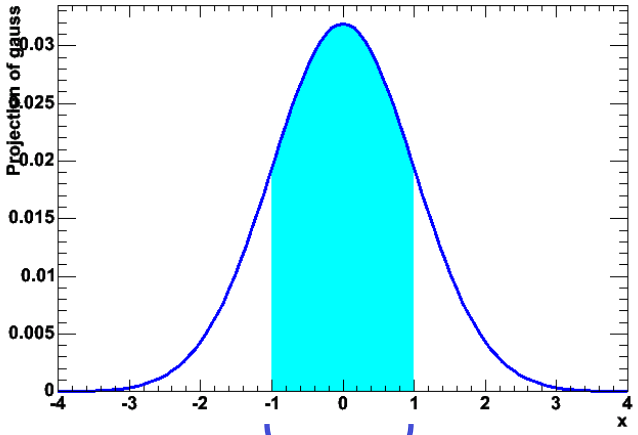
$$P(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

- *Mean* and *Variance*

$$\langle x \rangle = \int_{-\infty}^{+\infty} xP(x;\mu,\sigma)dx = \mu$$

$$V(x) = \int_{-\infty}^{+\infty} (x-\mu)^2 P(x;\mu,\sigma)dx = \sigma^2$$

$$\sigma = \sigma$$



- Integrals of Gaussian

68.27% within 1σ	90% → 1.645σ
95.43% within 2σ	95% → 1.96σ
99.73% within 3σ	99% → 2.58σ
	99.9% → 3.29σ

Measurement uncertainties

Measurement uncertainties

A measurement based on statistical data is meaningless without an estimate of the corresponding uncertainty

The measurement uncertainty is a measure of the expected deviation in the measured value with respect to the true value

if you're a frequentist…

Uncertainties (or "errors") come in two flavors:

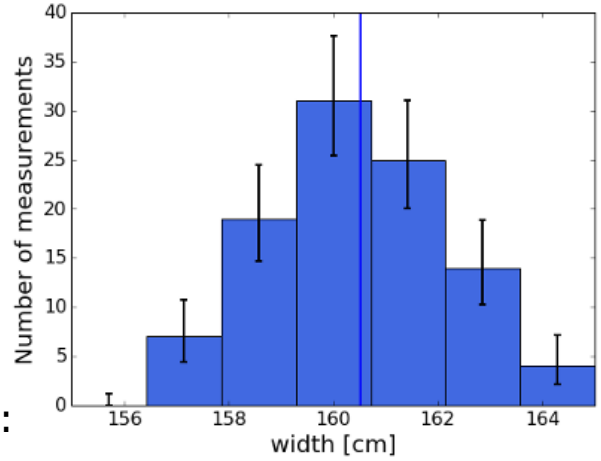
- Statistical uncertainties
  - Directly related to size of dataset (number of records)
  - Differences between measurements statistically equivalent datasets
- Systematic uncertainties
  - Not (directly) related to size of dataset (number of records)
  - Limitations in modelling/obtaining/processing common to all records

- Variety of methods to quantify uncertainties
- Only in trivial cases different methods agree exactly
- Up to the analyst to obtain a meaningful estimate

Measurement uncertainties

Example: size of storage cabinet

You are about to buy a storage cabinet at a furniture store, but you first want to make sure it fits into your bedroom. The critical dimension is the width, so you have asked 100 people in the store to measure this quantity separately on the cabinet displayed there, using your tape measure. You take the average of their measurements as your best estimate.

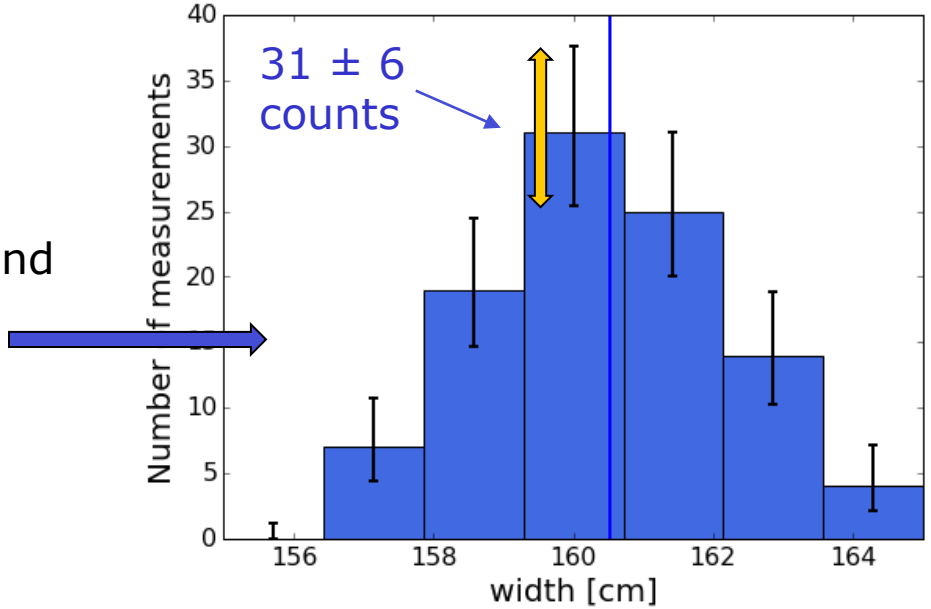


- Statistical uncertainty
  - Different deviation each measurement (std. σ_m):
    - Positioning of the tape measure
    - Reading the tape measure (between marks)
    - …
  - Decreases for large number of people ()
  - Uncertainty (std.) in average estimated by: $\sigma = \frac{\sigma_m}{\sqrt{N}}$
- Systematic uncertainty
  - Common deviation for all measurements
    - Tolerance of tape measure (calibration, temperature, …)
    - Width of displayed cabinet different from width of cabinet you buy
    - …
  - Most syst. sources do not depend on number of measurements

Measurement uncertainties - counting

Counting examples

- Golden cars in next hour
- Radioactive decays per second
- Measurements in width bin



31 counts in 160 cm bin

- How would this number vary when "experiment" is repeated?
- Recall Poisson distribution
  - Observed count: r
  - Expected count (infinitely many experiments): λ
  - For a collection of experiments
    - mean: λ estimated by $r=31$
    - standard deviation: $\sqrt{\lambda}$ estimated by $\sqrt{r} \approx 6$

$$P(r;\lambda) = \frac{e^{-\lambda}\lambda^{r}}{r!}$$

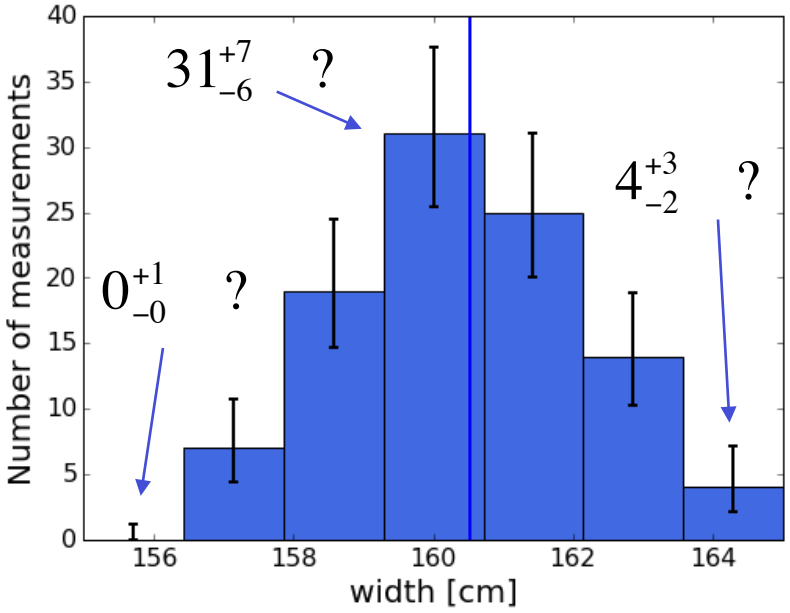
Note: total number of measurements is fixed in this experiment (N = 100); we actually need the binomial distribution here…
Out of 100 measurements, how many fall into this bin?

Measurement uncertainties - counting

Recall:

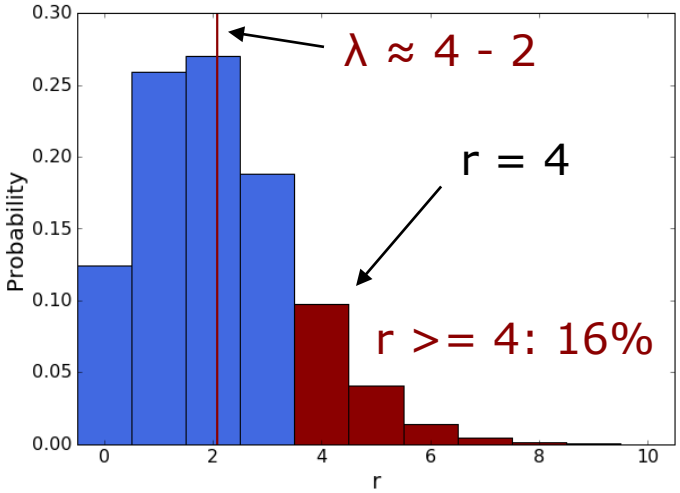
- Poisson → Gaussian for "large" λ (> 20)
  - 68% of "experiments" within μ ± σ = λ ± √λ
  - σ ≈ √λ is a sensible measure of uncertainty
  - r is a sensible estimate of λ

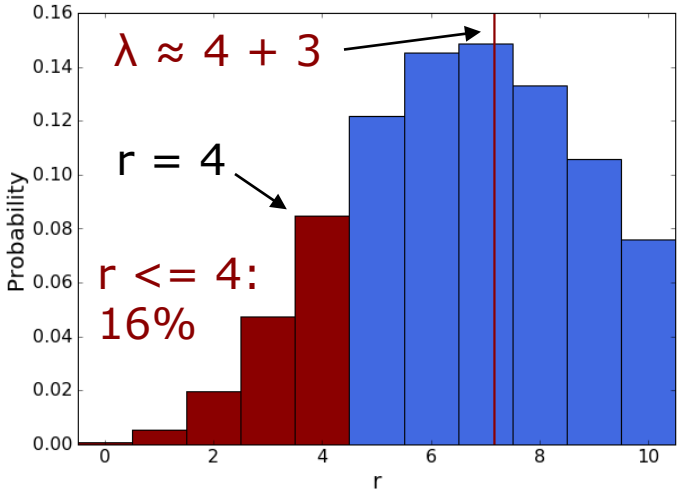
- Distribution *asymmetric* for small λ



Measurement uncertainties - counting

- Distribution *asymmetric* for small λ
  - Estimate λ interval instead:





$$P(r;\lambda) = \frac{e^{-\lambda}\lambda^{r}}{r!}$$

Values of λ with measured r or r further away from λ in only 16% of experiments

See also lecture on hypothesis testing

Measurement uncertainties

- Doing an experiment → making measurements (X_i)
- Measurements not perfect → imperfection quantified in the *resolution* or *error* or *measurement uncertainty* (σ_i)
  - Notation: $X_i \pm \sigma_i$
- Common language to quote errors
  - Gaussian standard deviation = sqrt(V(x))
  - 68% probability that true values is within quoted errors

  [NB: 68% interpretation relies strictly on Gaussian sampling distribution, which is not always the case, more on this later]
- *Central Limit Theorem:*
  Errors are usually Gaussian if they quantify a result that is based on many independent measurements

Central Limit Theorem

The Gaussian as 'Normal distribution'

- *Why are errors usually Gaussian?*

- The ***Central Limit Theorem*** says
  - If you take the sum X of N independent measurements x_i, each taken from a distribution of mean m_i, a variance $V_i=\sigma_i^2$, the distribution for x

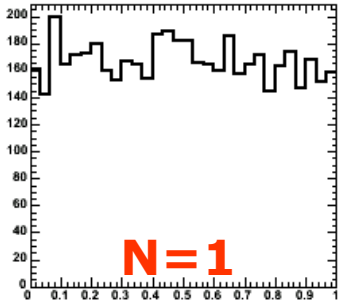
    (a) has expectation value $\langle X \rangle = \sum_i \mu_i$

    (b) has variance $V(X) = \sum_i V_i = \sum_i \sigma_i^2$

    (c) becomes Gaussian as N → ∞

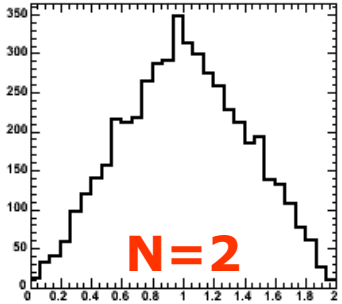
  - *Small print: tails converge very slowly in CLT, be careful in assuming Gaussian shape beyond 2σ*

Demonstration of Central Limit Theorem



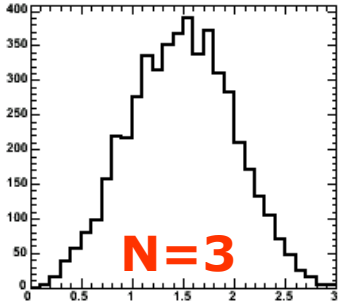
← 5000 numbers taken at random from a uniform distribution between [0,1].

- Mean = $^1/_2$, Variance = $^1/_{12}$

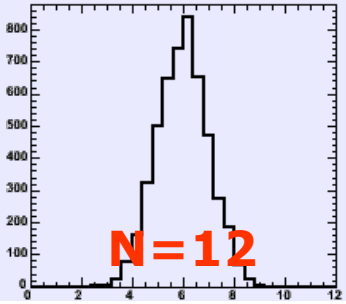


← 5000 numbers, each the sum of 2 random numbers, i.e. $X = x_1 + x_2$.

- Triangular shape



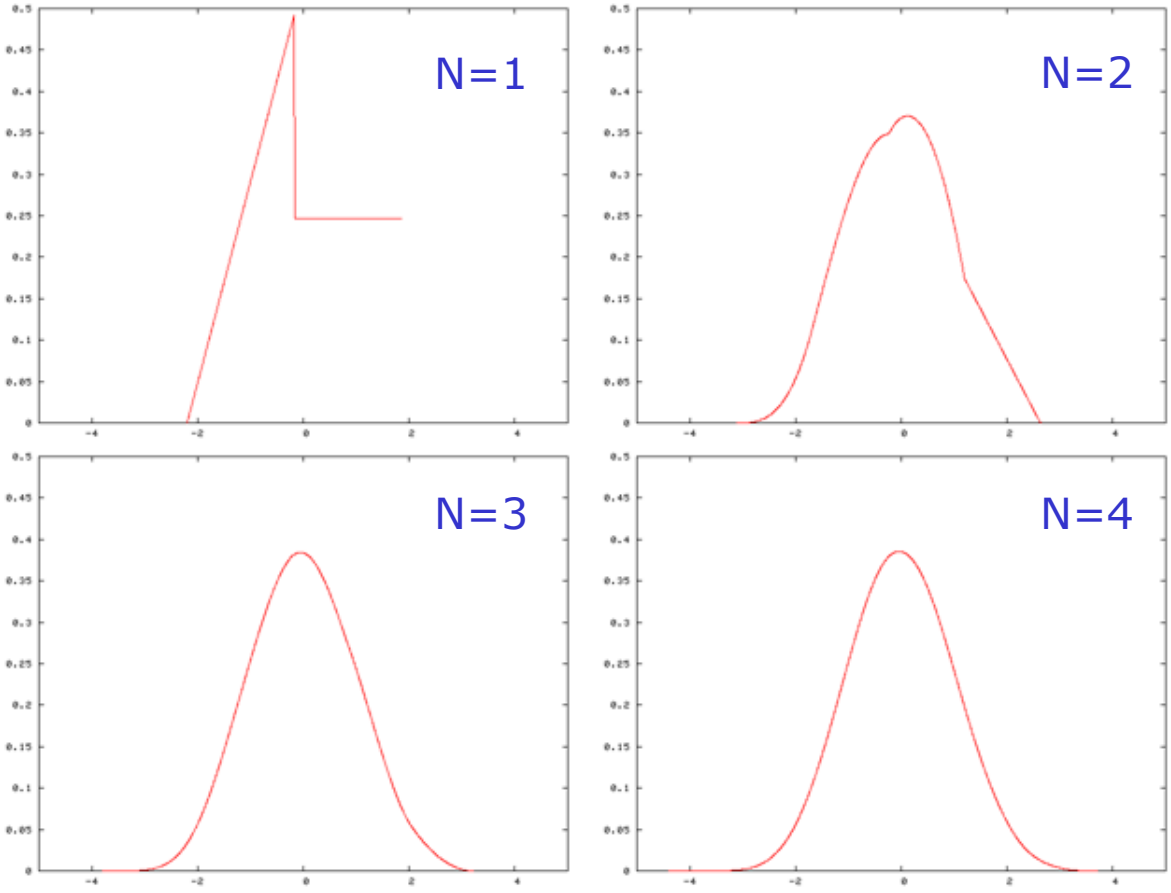
← Same for 3 numbers,
$X = x_1 + x_2 + x_3$



← Same for 12 numbers, distribution is (almost) exact Gaussian distribution

Another example of CLT

1. $X = x_1$
2. $X = x_1 + x_2$
3. $X = x_1 + x_2 + x_3$
4. $X = x_1 + x_2 + x_3 + x_4$



- From wikipedia: https://en.wikipedia.org/wiki/Central_limit_theorem

Central Limit Theorem – repeated measurements

- Common case 1 : Repeated identical measurements
  i.e. $\mu_i = \mu, \sigma_i = \sigma$ for all *i*

C.L.T

$$\langle X \rangle = \sum_i \mu_i = N\mu \quad \Rightarrow \quad \langle \bar{x} \rangle = \frac{X}{N} = \mu$$

$$V(\bar{x}) = \sum_i V_i(\bar{x}) = \frac{1}{N^2} \sum_i V_i(X) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$$

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{N}}$$ ← **Famous sqrt(N) law**

Central Limit Theorem – repeated measurements

- Common case 2 : Repeated measurements with identical means but different errors (i.e weighted measurements, $\mu_i = \mu$)

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

Weighted average

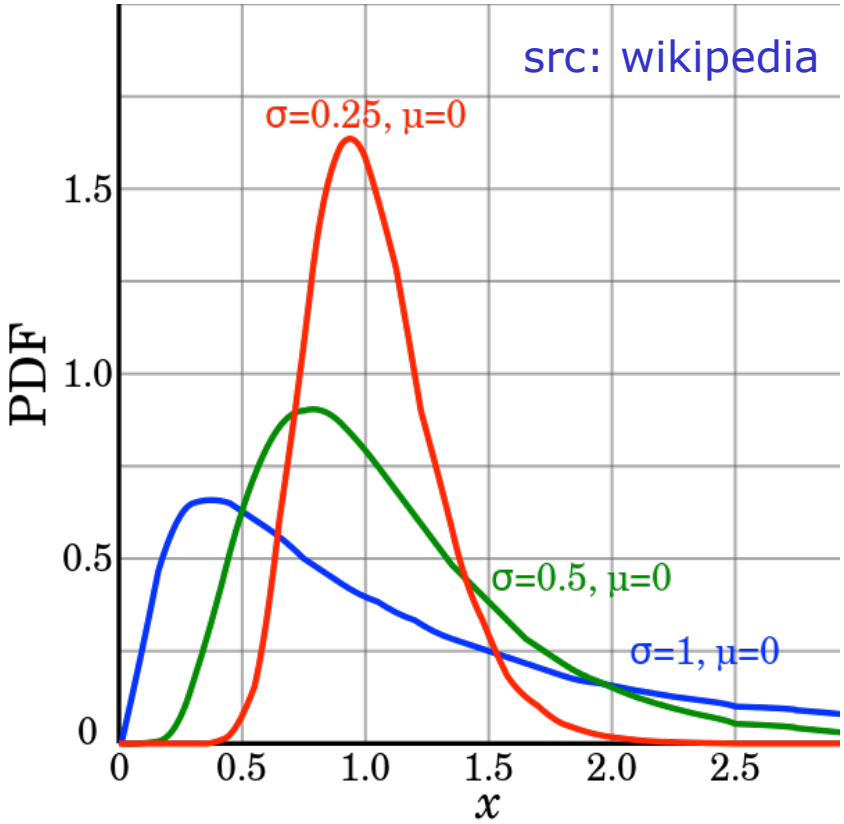
$$V(\bar{x}) = \frac{1}{\sum 1/\sigma_i^2} \Rightarrow \sigma(\bar{x}) = \frac{1}{\sqrt{\sum 1/\sigma_i^2}}$$

'Sum-of-weights' formula for error on weighted measurements

Central Limit Theorem – product of numbers

- For a *product* of arbitrary numbers (x>0), the central limit theorem results in a: log-normal distribution.

- E.g. length measurements.



- $$\ln \mathcal{N}(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right], \quad x > 0$$