Advanced Statistical Methods

Lecture 1

Homework and Exam

Homework assignments

- 2 x 2 hours TA sessions per week (Tuesday & Thursday 11-13h, same room)
- Homework is handed out at beginning of TA session and should be handed in one week later at end of TA session
- Help on the homework is provided during TA sessions
- Exercises require analytic work as well as numerical work on the computer
- Homework can be hand in hand-written, or send via Email (PDF)
- For numerical work, programs should be written as **Ipython Notebooks** and send via Email. They should "run out of the box" to give full points.

Exam

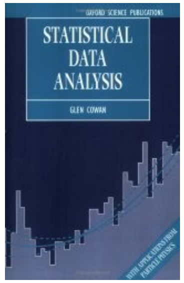
- There will be a written exam in the last session, on Thursday 29th January
- The total grade depends on both homework assignments (60%) as well as the exam (40%)

Contact

- Christoph Weniger (c.weniger@uva.nl)
- Michael Feyereisen (m.r.feyereisen@uva.nl)
- Richard Bartels (richard.t.bartels@gmail.com)

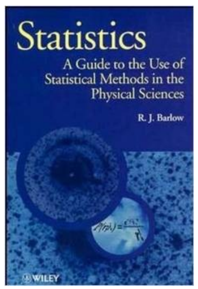
Slides & homework: https://staff.fnwi.uva.nl/c.weniger/
Later: Blackboard

Recommended Literature



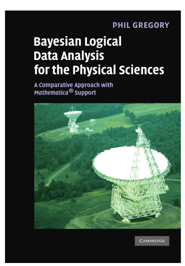
Glen Cowan, *Statistical Data Analysis,*
Oxford Science Publications, 1998

- Frequentist analysis, well known in Particle Physics
- Bonus: Monte Carlo methods and Unfolding



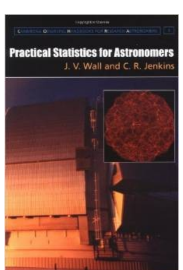
R. J. Barlow, *Statistics, A guide to the Use of Statistical Methods in the Physical Sciences,*
The Manchester Physics Series, 1988

- Traditional and very good book on Frequentist data analysis



P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences,* Cambridge University Press, 2005

- Bayesian "Bible", Conceptual introduction
- Many examples



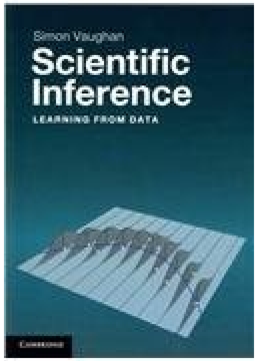
J.V. Wall and C.R. Jenkins, *Practical Statistics for Astronomers,* Cambridge Observing Handbooks for Research Astronomers, 2003

- Practical book for data analysis in Astronomy (both Frequentist and Bayesian)
- Many examples

Connection to Phil's Course

- Since most of you attended the course by Phil Uttley on Statistical Methods in Astrophysics and Astronomy (SM), I will assume that you know many of the basics, and continue from there.

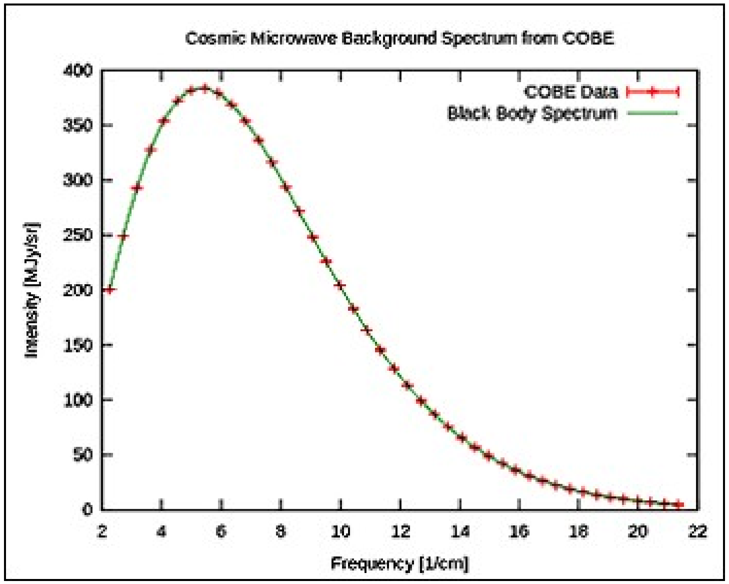
- SM was based on Simon Vaughan's book *Scientific Inference, Learning from Data.*



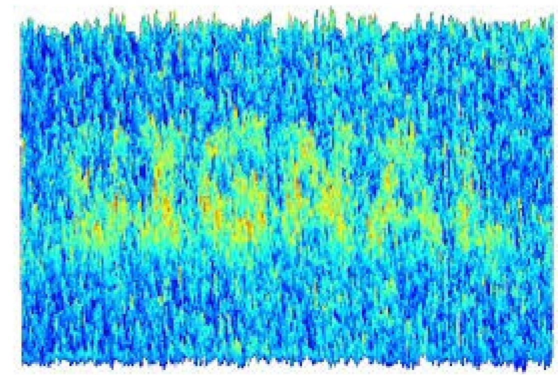
- In the first week of the present course, we will briefly repeat some of the most relevant material from SM as a reminder and to provide context for the rest of the course.

Understanding statistical tools matters

This course is about *why* statistical methods work.



- In cases where the experimental result is clear, the details of the statistical method often do not matter.
- In many cases, it is enough to apply standard statistical recipes (normal distribution, error propagation), to get reasonable results.



- When describing weak signals, close to the experimental threshold, *the details of the statistical method are crucial*
- Assumptions underlying the standard recipes might be violated
- It is important to understand not only who, but *why* statistical inference works. This is what the present course aims to do.

Overview

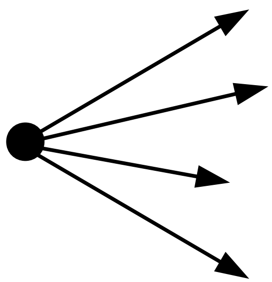
First four lectures

- Introduction: Bayesian and Frequentist statistics
- Probability distribution functions & Central limit theorem
- Frequentist analyses
  - Hypothesis testing
  - Estimators
  - Confidence intervals & Wilk's theorem
  - Profile likelihood technique & pitfalls
  - Trial factors & Coverage
  - Numerical minimizers
- Bayesian analyses
  - Basics: Evidence, Model selection, Credible intervals
  - Priors: Flat prior, Jeffery's prior, Non-informative priors
  - Sampling techniques: Markov Chain Monte Carlo, Multinest
- Applications & Advanced material
  - Principal component analysis
  - Angular power spectrum
  - Bootstrapping and Jackknife
  - ...

The two grand schools of statistical analysis

Frequentist

Fisher



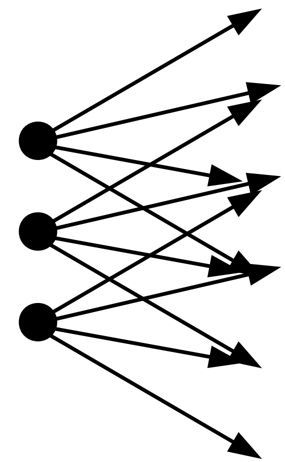
Deductive logic

- Based on “frequencies” of phenomena
- Central quantity: “p-value”

“Given a cause, what is the frequency (in repeated experiments) of a certain phenomenon to occur?”

Bayesian

Bayes



Inductive logic

- Probabilistic extension of logic
- “Posterior distribution”

“How does an observed phenomenon change my believe in different possible causes?”

Probabilities in a nutshell

"Probabilities" mean here

- Frequencies of events (in 1/6 of the cases the dice shows a six)
- Plausibility or believe in a proposition (the believe in "The Higgs boson exists.")

The most relevant rules

- Degrees of plausibility are represented by real numbers between 0 (not realized) and 1 (realized)
- Probabilities for *mutually exclusive* and *exhaustive* elementary events/propositions sum to one:

$$\sum_i P(X_i|I) = 1$$

(*I indicates b*ackground information)

- An event/proposition is either true or false (inference in binary logic)

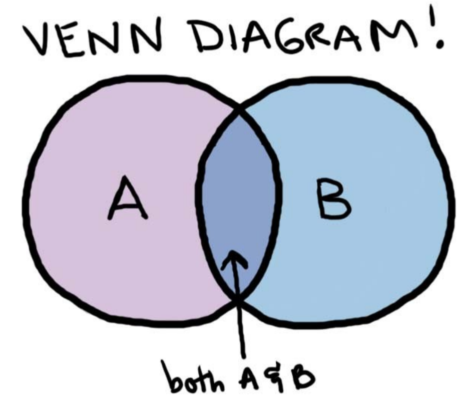
$$P(X|I) + P(\bar{X}|I) = 1$$

- Structural consistency (the result does not depend on the way of reasoning) is guaranteed by the rule for conditional probabilities

$$P(X, Y|I) = P(X|Y, I) \cdot P(Y|I)$$

- Elementary events/propositions follow the rules of set theory

$$P(A|I) + P(B|I) - P(A \cap B|I) = P(A + B|I)$$



(for a full discussion and derivation from fundamental requirements for consistent reasoning see Chapter 2.5 in Gregory)

Bayes' Theorem

Likelihood function

Posterior **Prior**

$$P(H|D,I) = \frac{P(D|H,I) \cdot P(H,I)}{P(D|I)}$$

Model evidence or **global likelihood**

It is a direct consequence of the rule for conditional probabilities:

$$P(X|Y,I) \cdot P(Y,I) = P(X,Y|I) = P(Y|X,I) \cdot P(X,I)$$

Notes:

- Bayes' theorem provides a rule for how to *update* the probability or plausibility of a certain hypothesis *H* to be true in light of data *D*. This always depends on additional background information *I*, which is often not made explicit.
- Frequentists are interested in likelihood functions *only*

$$\mathcal{L}(H|D,I) \propto P(D|H,I)$$

  It is in general *not* equal to the posterior, which is most obvious looking at the normalization of the functions (with x and θ being data and model parameters, respectively).

$$\int d\theta P(\theta|x) = 1$$

Typical Frequentist questions

Fisher

There is a new test for the *Schnitzler syndrome (c)* with the characteristics:

- 5% false positive $P(p|\bar{c}) = 0.05$
- 10% false negative $P(p|c) = 0.9$

You order the test online, and get a positive result.
Should you be worried?

The Frequentist says:
"I can exclude the null-hypothesis of not having the Schnitzler syndrome at 95% CL. End of story."

Caveat: There are **hidden trials**

- There might be 20 other people having done the test, none of them having the disease. Still, one of them will get a positive result on average.
- Maybe you also did tests for other diseases.
- Maybe *nobody* had the Schnitzler syndrom in the last 100 years

$\rightarrow$ Chances for you to having the disease could be still very low.

One could account for hidden trials by making abstract statements about the frequency of wrong and right statistical statements (instead of observations). This is exactly what Bayesian inference forces us to do from the start.

Typical Bayesian questions

There is a new test for the *Schnitzler syndrome* with the characteristics:

- 5% false positive $P(p|\bar{c}) = 0.05$
- 10% false negative $P(p|c) = 0.9$

You order the test online, and get a positive result. Should you be worried?

The Bayesian says:
"What are the priors?"

1:100000 persons have the disease

Bayes' theorem

$$P(c|p) = \frac{P(p|c)P(c)}{P(p)}$$

Prior
$$P(c) = 10^{-5}$$

Global likelihood
$$P(p) = P(p|c)P(c) + P(p|\bar{c})P(\bar{c}) \simeq 0.05$$

This yields a very low posterior probability:

$$P(c|p) \simeq \frac{P(p|c)}{P(p|\bar{c})}P(c) = \frac{0.9}{0.05}10^{-5} \simeq 2 \times 10^{-4}$$

Pros and Cons of the two approaches

	Frequentist	**Bayesian**
Pro	**Pro**: • No prior dependence (what is the prior for a flat Universe?) • Objective procedure • Clear interpretation of results	**Pro**: • Prior dependence is formalized • Reasoning about causes, not observations • Hard to use completely wrong (this is a conjecture to be tested in this course)

Con	**Con**:
	• Be aware of *hidden trials*
	  • Publication bias, many researches
	  • Many ways of calculating p-values
	• "Frequencies" refer to repeated experiments with *exactly the same conditions*
		Con:
		• Results are difficult to show in a prior-independent way
		• Some people think it is "esoteric"
		• Difficult for *non-parametric* studies

Basic definitions I

- A characteristic of a system is said to be **random** when it is not known or cannot be predicted with complete certainty.
- The degree of randomness can be quantified with the concept of **probability** (or frequencies; in the Frequentist sense).
- The **sample space** consists of a certain set of elements that are the values or properties that a random variable can acquire.

$$x \sim X \in S$$

- The probability distribution function describes the probability (either as frequency or subjective probability) that a certain value is realized.

$$P(x_i|H) \quad \text{Probability mass function (PMF)}$$

$$P(x_1 < x < x_2|H) = \int_{x_1}^{x_2} dx\, P(x|H) \quad \text{Probability density function (PDF)}$$

- In general, it depends on prior assumptions and hypothesis, here summarized as *H*.

Basic definitions II

- **Mean value** for discrete or continuous distributions

$$\langle x \rangle \equiv \frac{1}{n} \sum_{i=1,\ldots,n} x_i P(x_i|H) \qquad \langle x \rangle \equiv \int dx\ x\ P(x|H)$$

- **Variance** and **standard deviation**

$$var(x) = \langle (x - \langle x \rangle)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2 \quad \text{and} \quad \sigma = \sqrt{var(x)}$$

- **Covariance**

$$covar(x, y) = \langle xy \rangle - \langle x \rangle \langle y \rangle$$

- **Median** x_m

$$\int_{x_m}^{\infty} dx\ P(x|H) = \int_{-\infty}^{x_m} dx\ P(x|H) = 0.5$$

- **Mode**

$$x_{\text{mode}} = \arg\max_x P(x|H)$$

- **Skewness**

$$skew(x) = \langle (x - \langle x \rangle)^3 \rangle / \sigma^3$$

- **n-th central moment:**

$$\mu_n = \langle (x - \langle x \rangle)^n \rangle$$

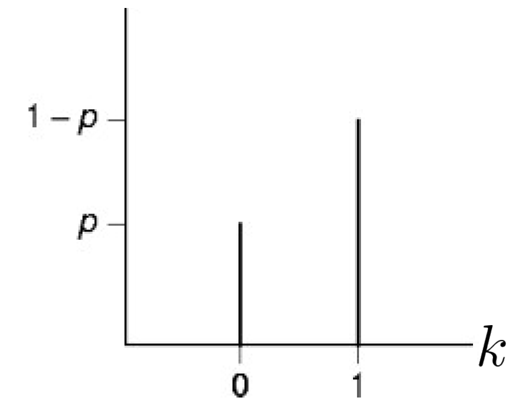
Important discrete distributions

Bernoulli distribution

A single yes/no question, answered yes (1) with probability *p*

$$P(k|p) = p^k (1-p)^{1-k} \qquad k \in \{0,1\}$$

$$\langle k \rangle = p$$
$$var(k) = p(1-p)$$



- *Example:* Throw a biased coin

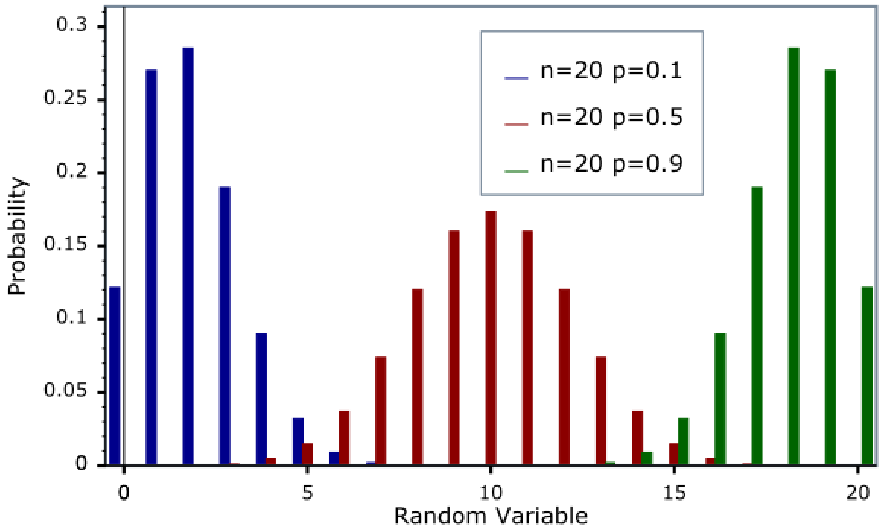
Binomial distribution

Number of successes in draw of *n* elements with individual success probability *p*.

$$P(k|n,p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

$$\langle k \rangle = np \qquad var(k) = np(1-p)$$

- *Example:* Draw of colored beans from a large bin



The Poisson Distribution

Poisson distribution

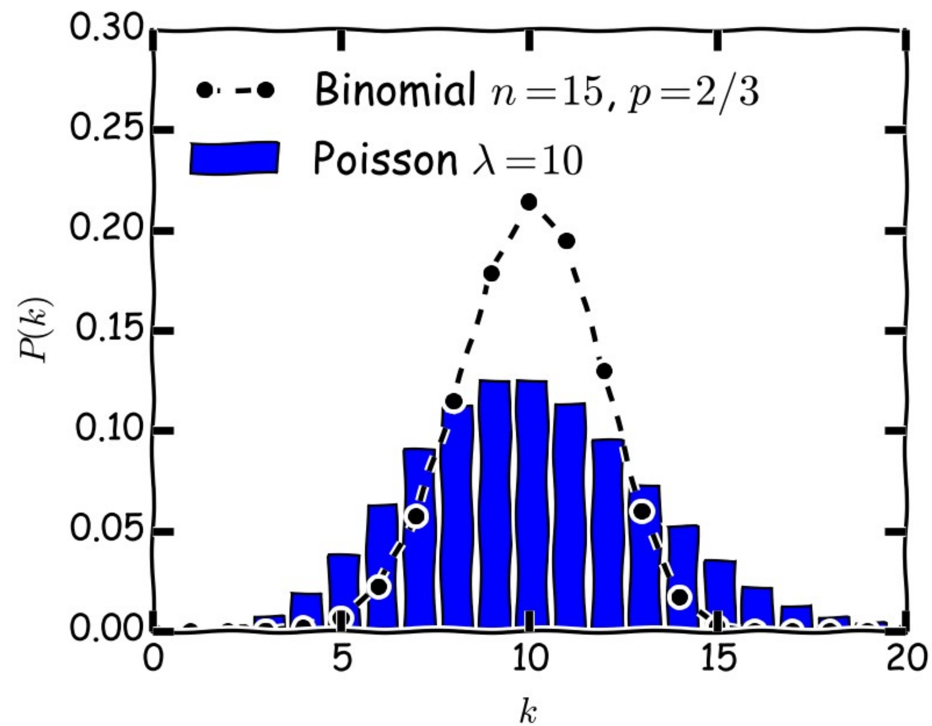
$$P(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad \langle k \rangle = \lambda \qquad var(k) = \lambda$$

- *Example*: Number of detected photons from an radioactive source
- *Note:* The sum of *N* Poisson-distributed random variables is Poisson distributed, with mean

$$k = \sum_i k_i \qquad \lambda = \sum_{i=1}^{N} \lambda_i$$

- Follows from Binomial distribution in the limit

$$n \to \infty, \text{ keeping } \lambda \equiv pn \text{ fixed}$$



Normal and chi-squared distribution

Gaussian / Bell curve / Laplacean / Normal distribution

$$P(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \langle x \rangle = \mu \qquad var(x) = \sigma \qquad x \sim N(\mu,\sigma)$$

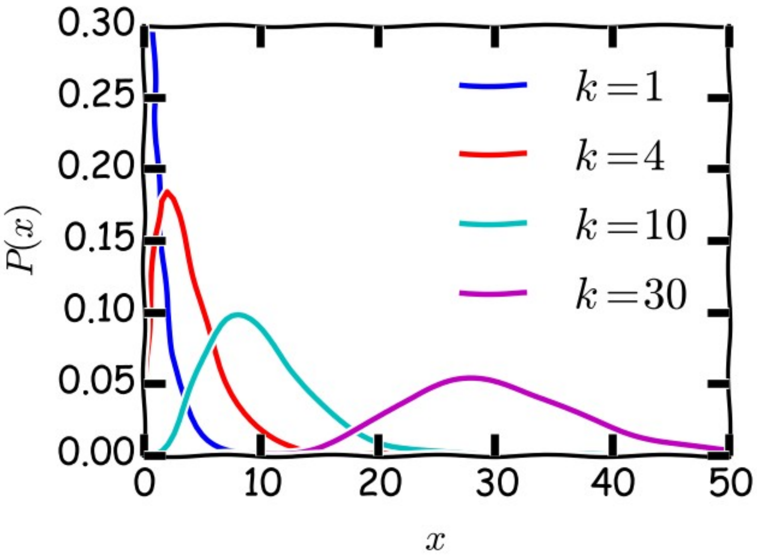
Notes:

- Its central importance comes from the *central limit theorem*
- Many random variables are normal distributed in practice, but the reasons for that are often complex.
- *"Everybody believes in the law of errors, the experiments because they think it is a mathematical theorem, the mathematicians because they think it is a experimental fact."* (Lippmann; Barlow p36)

Chi-squared distribution

Describes statistical distribution of outliers

$$P(x|k) = \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)} \qquad \langle x \rangle = k \qquad var(x) = 2k \qquad x \sim \chi^2_k$$



- Is defined as the sum of squares of normal distributed variables

$$x = x_1^2 + \cdots + x_k^2 \qquad x_i \sim N(\mu = 0, \sigma = 1)$$

- Describes the statistical distribution of *outliers*
- Important because of *Wilks' theorem*

Multivariate normal distribution

"Rotated" normal distributions

- The PDF is similar to the 1-dim case

$$P(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})\right)$$

with mean and variance

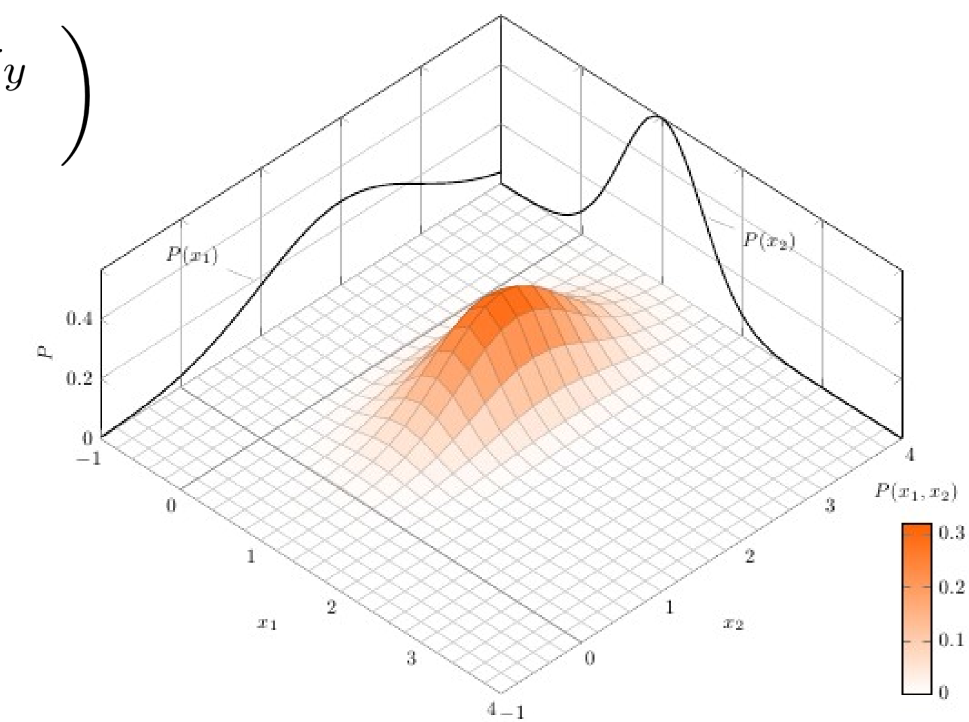
$$\langle \vec{x} \rangle = \vec{\mu} \qquad \langle (x_i - \mu_i)(x_j - \mu_j) \rangle = \Sigma_{ij}$$

- The two-dimensional case with variables *x*, *y*

$$\vec{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

where ρ is the *correlation* between x and y.

- Note that each x_i individually is normal distributed with variance σ_i.



Other useful distributions

Exponential distribution

$$P(x) = \frac{1}{\xi} e^{-\xi x} \quad (x \geq 0)$$

Uniform distribution

$$P(x) = \frac{1}{\beta - \alpha} \quad (\beta \geq x \geq \alpha)$$

Log-normal distribution

- *ln(x)* is normal distributed

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{x} \exp\left(\frac{-(\ln x - \mu)^2}{2\sigma^2} \right)$$

Cauchy distribution / Breit-Wigner distribution

- The proper Cauchy distribution is obtained for $x_0 = 0, \quad \Gamma = 2$

$$P(x) = \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x - x_0)^2}$$

- Though this distribution is omnipresent in particle physics, its convergence behavior is extremely bad

Student's t-distribution

- Generalization of unit normal distribution when variance is estimated from data

$$P(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)} \frac{1}{(1 + (x^2/n))^{(n+1)/2}}$$

Many interrelations

Sum rules

$$x_1 + x_2 = x$$

- Poisson + Poisson = Poisson
- Normal + Normal = Normal
- Chi-squared + chi-squared = chi-squared
- Cauchy + Cauchy = Cauchy

Important approximations

- Binomial to Poisson

$$\lambda \underset{n \to \infty}{=} pn$$

- Poisson to Normal

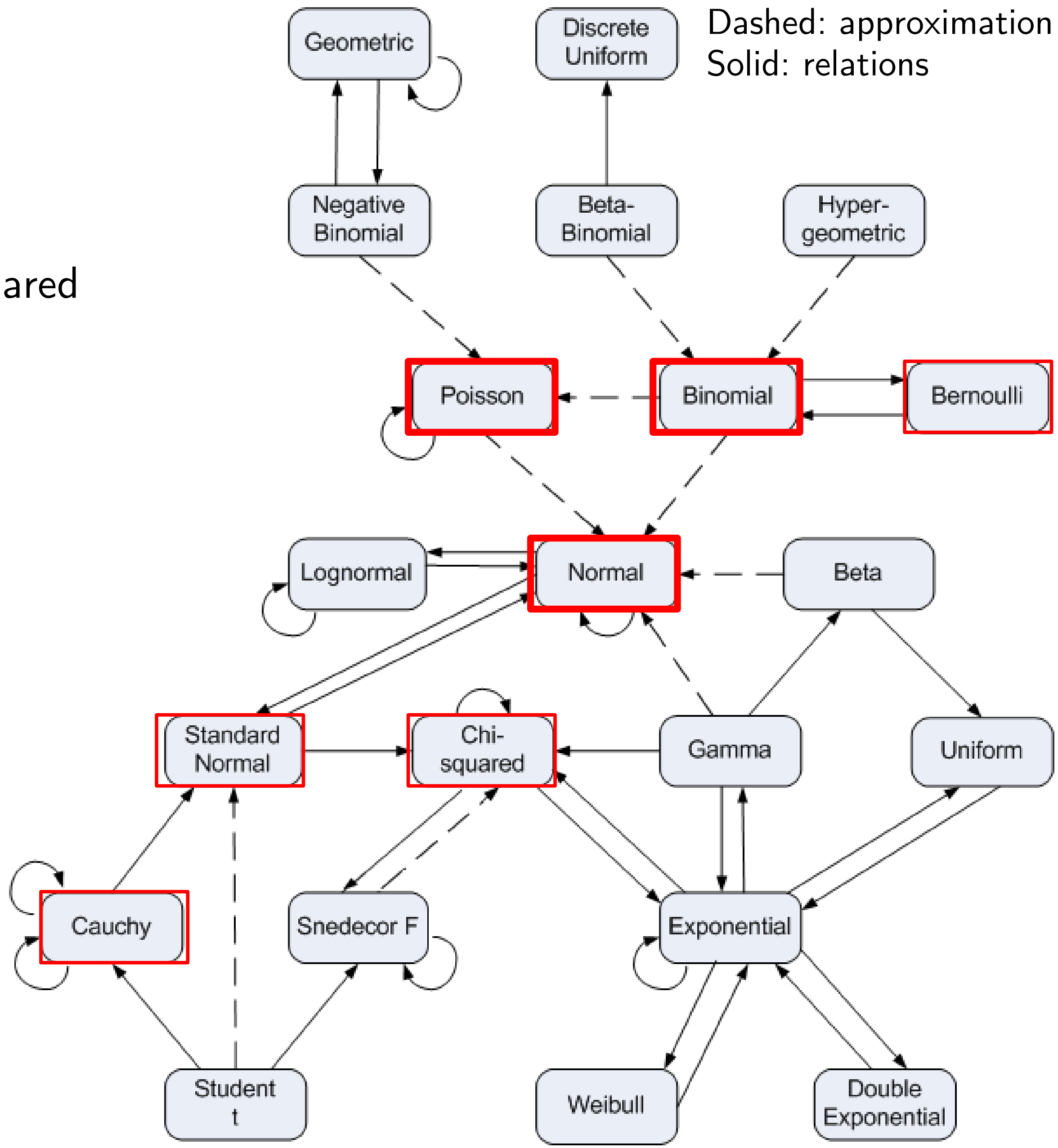
$$\mu,\ \sigma \underset{\lambda \to \infty}{=} \lambda$$

- Chi-squared to Normal

$$\mu,\ \frac{\sigma}{2} \underset{k \to \infty}{=} k$$

- Student t to Standard Normal

$$n \to \infty$$



From http://www.johndcook.com/blog/distribution_chart/

The central role of the normal distribution

The Central Limit Theorem (CLT)

The sum of *n* independent continuous random variables,

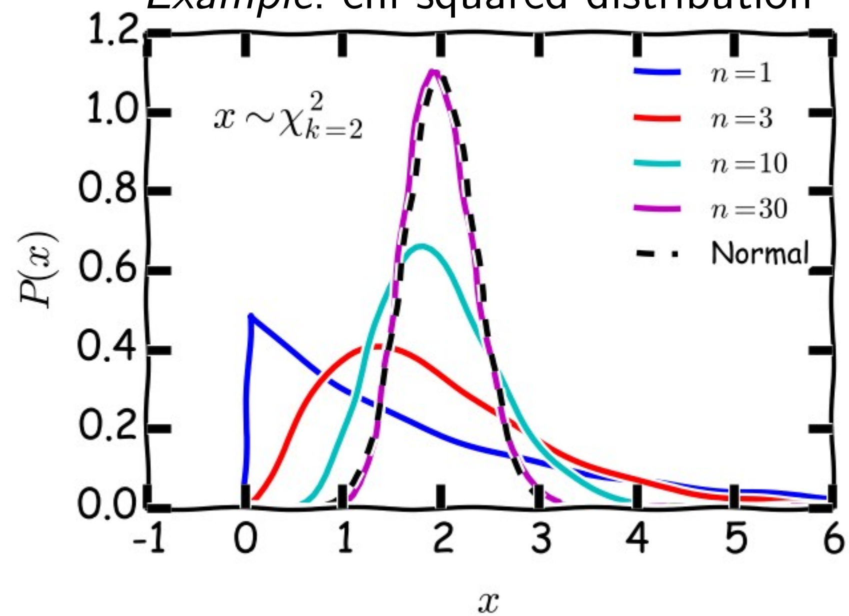
$$x = \frac{1}{n}(x_1 + \cdots + x_n)$$

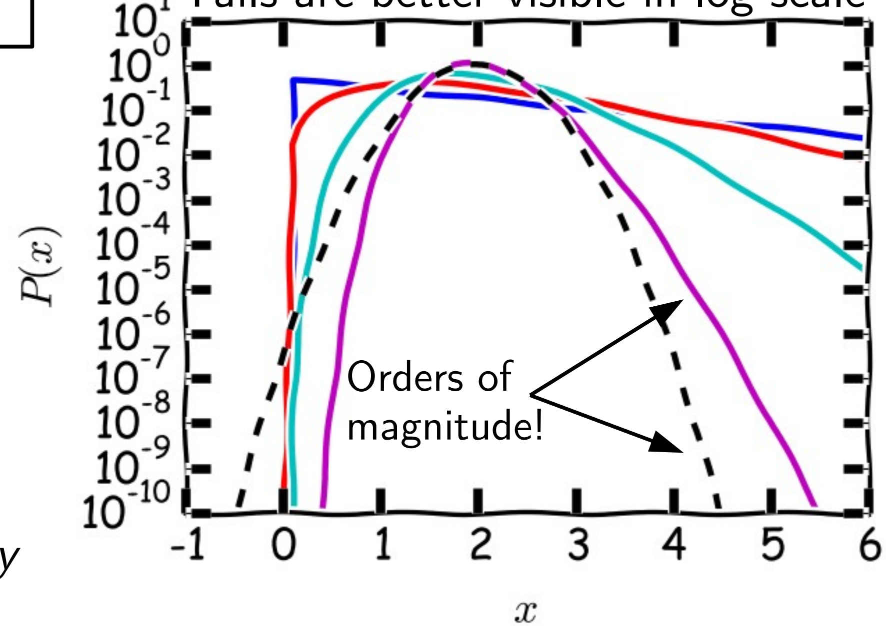
with means and variances

$$\text{mean}: \ \mu_i \qquad \text{variance}: \ \sigma_i^2$$

becomes a Gaussian random variable with mean and variance

$$\langle x \rangle = \frac{1}{n}\sum \mu_i \qquad var(x) = \frac{1}{n^2}\sum \sigma_i^2$$





Notes:

- The CTL holds for a very large number of underlying distributions, but for some it completely fails.
- In general, the CTL works better at the center of the distribution than far away from the center.

Warning:

- Even if the center of the summed distribution is indistinguishable from a Gaussian, there might be *very large deviations in the "wings" or "tails"!*

Proof of the Central Limit Theorem

We are interested in the following PDF (the sum of variables with distributions *f*, *g*, *h*):

$$P(x) = \int dx' dx'' f(x - x') g(x' - x'') h(x'')$$

A few useful definitions:

A) The Characteristic Function.

$$\tilde{P}(k) = \langle e^{ikx} \rangle = \int e^{ikx} P(x) dx$$

Note: Convolutions simplify to multiplications

$$\tilde{P}(k) = \tilde{f}(k) \tilde{g}(k) \tilde{h}(k) \ldots$$

B) Cumulants. Define by the Taylor expansion of the log of the characteristic function

$$\ln \tilde{P}(k) = (ik)\kappa_1 + \frac{(ik)^2}{2!}\kappa_2 + \ldots$$

$$\kappa_i = \kappa_i^f + \kappa_i^g + \ldots$$

Proof of the Central Limit Theorem

The first three cumulants are functions of the mean, the variance and the skew

$$\kappa_1 = \langle x \rangle$$ mean

$$\kappa_2 = \langle x^2 \rangle - \langle x \rangle^2$$ variance

$$\kappa_3 = \langle x^3 \rangle - 3\langle x \rangle \langle x^2 \rangle + 2\langle x \rangle^3$$

For normal distribution: $\kappa_{i \geq 3} = 0$

Adding *N* distributions with cumulant $\bar{\kappa}_r \approx \kappa_r^f, \kappa_r^g, \ldots$ implies

(notational simplifications)

$$\kappa_r = N\bar{\kappa}_r$$

To see what this means, we rescale *x* such that the variance equals one

$$x \rightarrow \frac{x}{\sqrt{N\bar{\kappa}_2}}$$

This implies

$$\kappa_r \rightarrow \frac{N\bar{\kappa}_r}{(N\bar{\kappa}_2)^{r/2}}$$

and in the large *N* limit

$$\underset{r \geq 3}{\overset{N \to \infty}{\longrightarrow}} 0$$

Hence, for large enough values of *N*, only the first two cumulants are important.