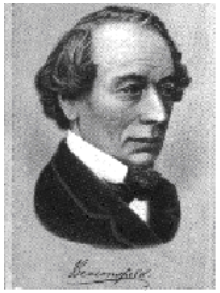
Statistics in astronomy (I) introduction

Shiyin Shen

Generally, statistics has got a bad reputation

Mark Twain



Benjamin Disraeli

"There are three types of lies: lies, damned lies and statistics"

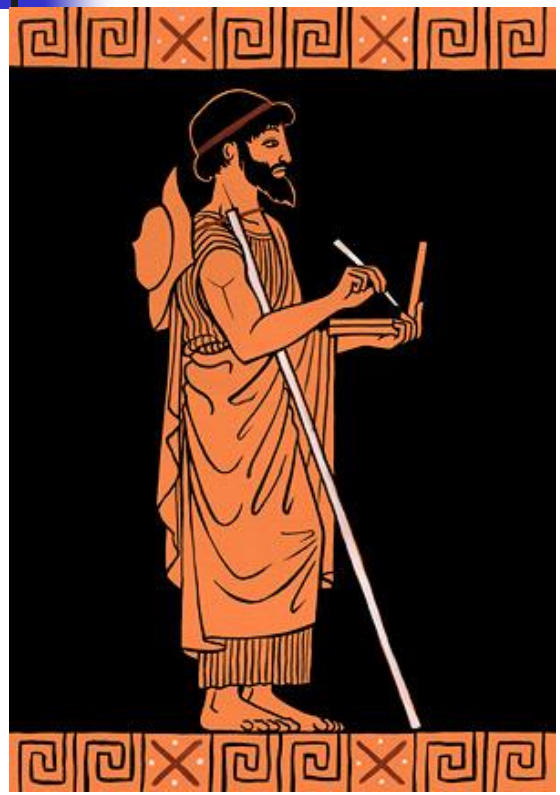
Often for good reason:

The Economist

Jun 3rd 2004

… two researchers at the University of Girona in Spain, have found that 38% of a sample of papers in *Nature* contained one or more statistical errors…

Right-thinking gentlemen #1



Herodotus, c.500 BC

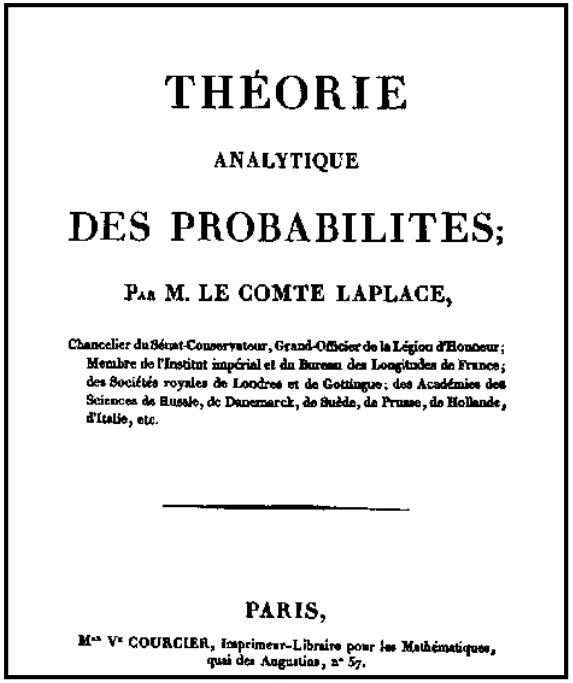
"A decision was wise, even though it led to disastrous consequences, if with the evidence at hand indicated it was the best one to make; and a decision was foolish, even though it led to the happiest possible consequences, if it was unreasonable to expect those consequences"

We should do the best with what we have, not what we wished we had.

Right-thinking gentlemen #2

Pierre-Simon Laplace
(1749 – 1827)

"Probability theory is nothing but common sense reduced to calculation"



THÉORIE

ANALYTIQUE

DES PROBABILITES;

PAR M. LE COMTE LAPLACE,

Chancelier du Sénat-Conservateur, Grand-Officier de la Légion d'Honneur; Membre de l'Institut impérial et du Bureau des Longitudes de France; des Sociétés royales de Londres et de Gottingue; des Académies des Sciences de Russie, de Danemarck, de Suède, de Prusse, de Hollande, d'Italie, etc.

PARIS,

Mme Ve COURCIER, Imprimeur-Libraire pour les Mathématiques, quai des Augustins, n° 57.

1812.

Right-thinking gentlemen #3

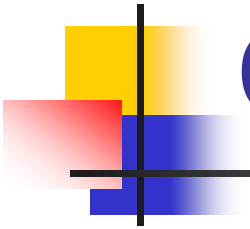
Occam's Razor

William of Occam
(1288 – 1348 AD)

"Frustra fit per plura, quod fieri potest per pauciora."

"It is vain to do with more what can be done with less."

Everything else being equal, we favour models which are ***simple***.

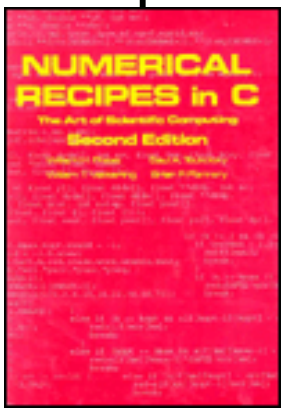


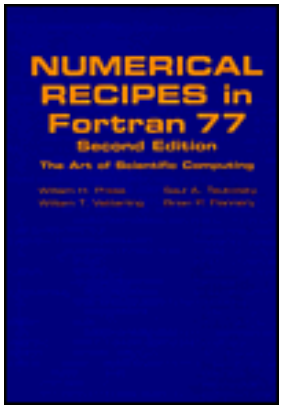
Contents

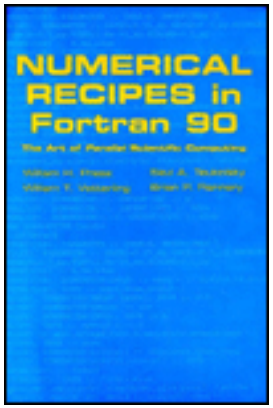
- Probability and statistics
  - Probability distribution function
    - Poisson, Gaussian, chi-square
- Modeling of data
  - statistical tests
    - χ^2, t, F-test
  - Parameter estimation
    - minimum χ^2
    - Maximum likelihood
  - Non-parametric statistics
    - K-S test, spearman rank correlation

References

- Numeric recipes
  - Chapter 14 and 15







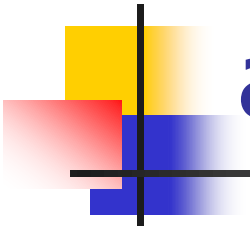
http://www.numerical-recipes.com/

- Astronomical statistics (Andy Taylor 2004)

Why statistics in astronomy?

- Joke of the white sheep on the grassland
- Cosmology principle
- Only observation, no experiment
  - e.g. the viewing angle of a galaxy
- Individual VS common properties
  - e.g. AGN unified model
- Big uncertainties in observation
- Data mining
  - Huge mount of data accumulated from morden surveys

What statistics can do in astronomy



- Detection of signals
  - source detection, spectral features
  - Correlations: significant?
- Modeling data
  - Is our sample ‘fair’?
  - How data confirm or rule out a theory?
  - If a model supposed to be right, how to estimate the model parameters?

How often do astronomers need statistics?

Of ~15,000 refereed papers annually:

1% have `*statistics*' in title or keywords
5% have `*statistics*' in abstract
10% treat variable objects
5-10% (est) analyze data tables
5-10% (est) fit parametric models

What is probability

- Frequency: The probability of the prize of a lottery.
- Lack of information: e.g. the probability of tomorrow raining
- Q: Shall the insurance company refund your premium if no accident happens

Two approaches

- If we measured the mean mass of a sample of G stars. What the meaning if we say that at the 68% confidence level the mean mass of G star is $a \pm b$
- Frequentist (classical): if M=a, we would expect a sample mean in the range $a \pm b$ for 68% of the times
- Bayesian: the true mean mass M of G stars lies in the range $a \pm b$ has a 68% probability of being true

Frequentist VS Bayesian

- Frequentist
  - Data are random, probability is frequency of data
  - Cannot refer to the probability of a hypothesis (either true or false)
- Bayesian
  - Data are not random (in astronomy!)
  - Evaluate the probability of a hypothesis in light of data (and prior information)
- Different philosophically, but agree on each other in basic cases
  - Reference: <The promise of Bayesian inference for astrophysics> by Loredo

Probability distribution function (PDF)

- s P(x) dx =1
- P either x or y = P(x)+P(y)
- P both x and y = P(x)P(y)
- Probability conservation: P(x)=g(y(x))|dy/dx|
- Moments of PDF

$$\left\langle x^n \right\rangle = \int_{-\infty}^{\infty} x^n \, p(x) dx$$

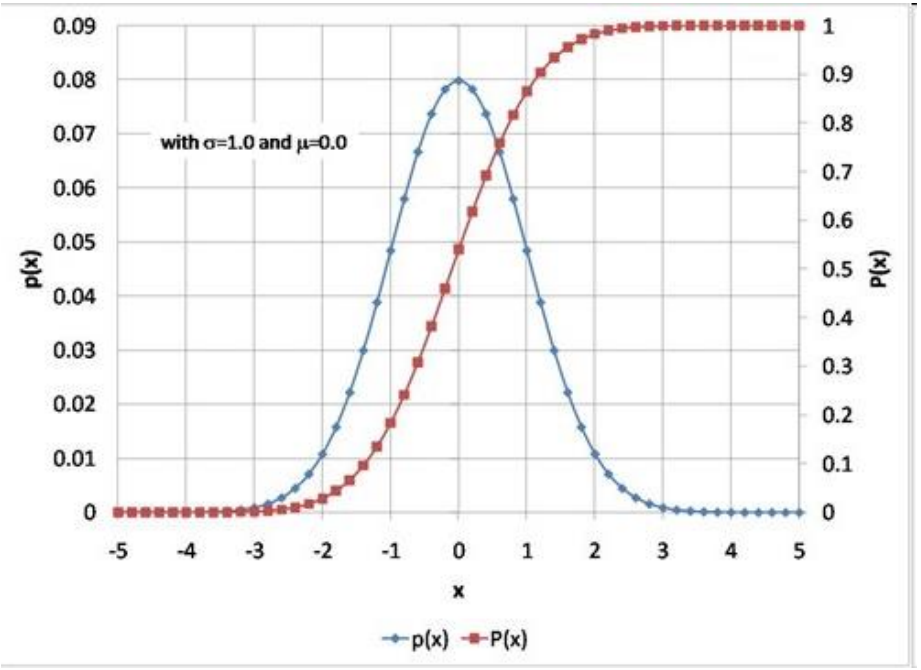
  - Mean μ =s x P(x) dx
  - Variance V = s $(x-\mu)^2$ P(x) dx

Cumulative distribution function

- Monotonic function between 0 – 1 from min to max

- For Gaussian x(0, 1), the probability of -1<x<1 is 68%

$$\Pr ob(x < a) = \int_{-\infty}^{a} p(x)dx$$



Measurement: mean, variance

- Sample: random realization of a distribution
- Sample mean $\bar{x} = \frac{1}{n}\sum x_i$
  - Sample mean is the unbiased estimator of mean of distribution
  - Variance of sample mean: 1/N of distribution Variance

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \langle x \rangle)^2,$$

- Sample variance
  - $N/(N-1)S^2$ is the unbiased estimator of variance of distribution

Mean VS median

- Median $\mathrm{P}(X \leq m) = \mathrm{P}(X \geq m) = \int_{-\infty}^{m} f(x)\,\mathrm{d}x = \frac{1}{2}.$
  - Mean = Median if p(x) is asymmetric
- For Gaussian distribution (μ, σ^2)
  - The variance of sample mean is σ^2/N
  - the variance of sample median is $\pi/2\ \sigma^2/N$
- For distributions with long tails
  - The variance of sample median is smaller than mean
  - Corresponding to Gaussian, use 68% region as a estimation of dispersion

Variance

- Standard deviation σ: $V(x)=\sigma^2$
  - Root mean scatter (RMS)
  - Error of σ: $0.71\ \sigma/\sqrt{N}$
- Scatter, dispersion
  - σ is a common example of dispersion
- Error/uncertainty: difference between the measured or calculated value and a true one.
  - Error is typically Gaussian distributed, characterized by the variance
- For independent variables: $V(x_1+..+x_n)=\sum_{i=1,N} V(x_i)$

Error propagation

- z=x+y, x,y independent random variables

$$p(z) = \int_{-\infty}^{\infty} dy\, p(z-y)p(y)$$

  - Var(z)=Var(x)+Var(y)
  - $\sigma_{x+y}^2=\sigma_x^2+\sigma_y^2$
- z=f(x,y)

$$f(x,y) = f(x_0,y_0) + x\frac{\partial f}{\partial x} + y\frac{\partial f}{\partial y}.$$

$$\sigma_z^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2.$$

Basic possibility distributions

- Discrete distributions
  - Binominal distributions
  - Poisson distributions
- Continuous distributions
  - Gaussian distribution
  - χ^2 distribution

Binomial distribution

- Number of 'successes' from N observations, for two mutually exclusive outcomes ('Heads' and 'Tails')
  e.g. number of binary stars, Seyfert galaxies, supernovae…

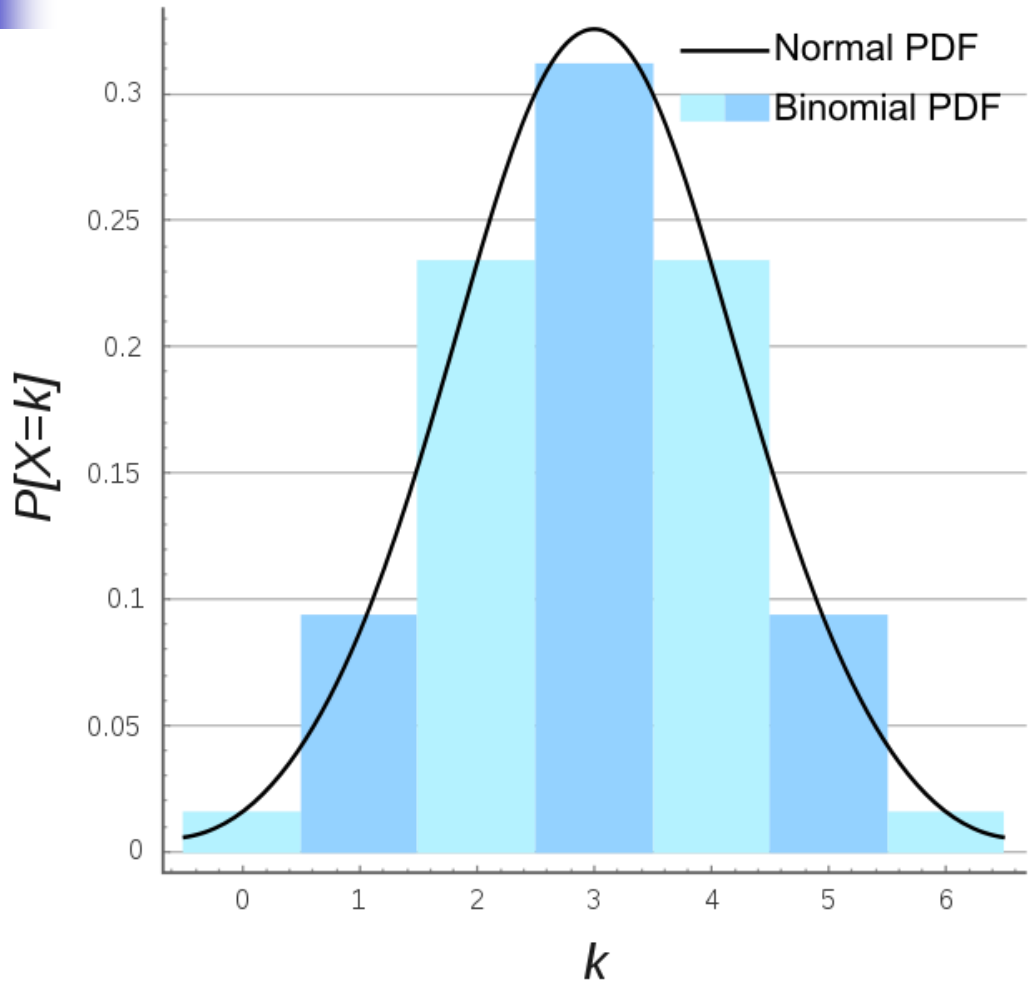
r = number of 'successes'

θ = probability of 'success' for single observation

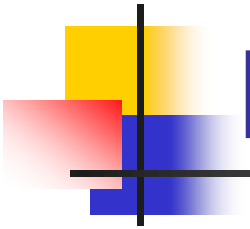
$$p_N(r) = \frac{N!}{r!(N-r)!}\theta^r(1-\theta)^{N-r}$$

Mean: Nθ Variance: Nθ(1-θ)

Example of Binomial distribution



N=6, p=0.5



practice

- Q1: What is the probability of heads come up 12 times when flip a coin 20 times?
  - Prior p=0.5
- Q2: Flip a coin 20 times, get 8 heads and 12 tails. What is the probability of heads come up?
  - data → model：Bayesian approach

Poisson distribution

- Data in discrete intervals, independent of each other, e.g. Energy channel of detector, location on the sky, time of arrival
  - probability of observing the counts n
    - μ: expectation value

$$P_n = \frac{\mu^n e^{-\mu}}{n!}$$

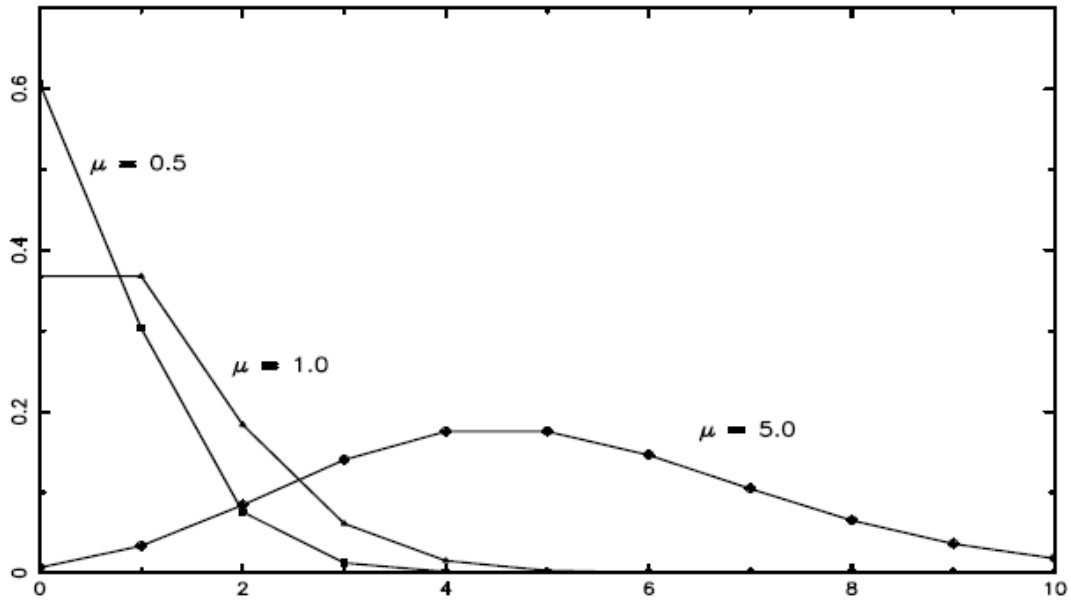
  - Limit case of Binominal distribution when N→1 and p→0
    - Number of photons we detected: a tiny fraction (p<<1) of photons emitted by star N>>1
    - Number of galaxies in a small piece of sky

Poisson distribution

Examples:

When we count N galaxies in a cell, we say the error of N is $N^{0.5}$

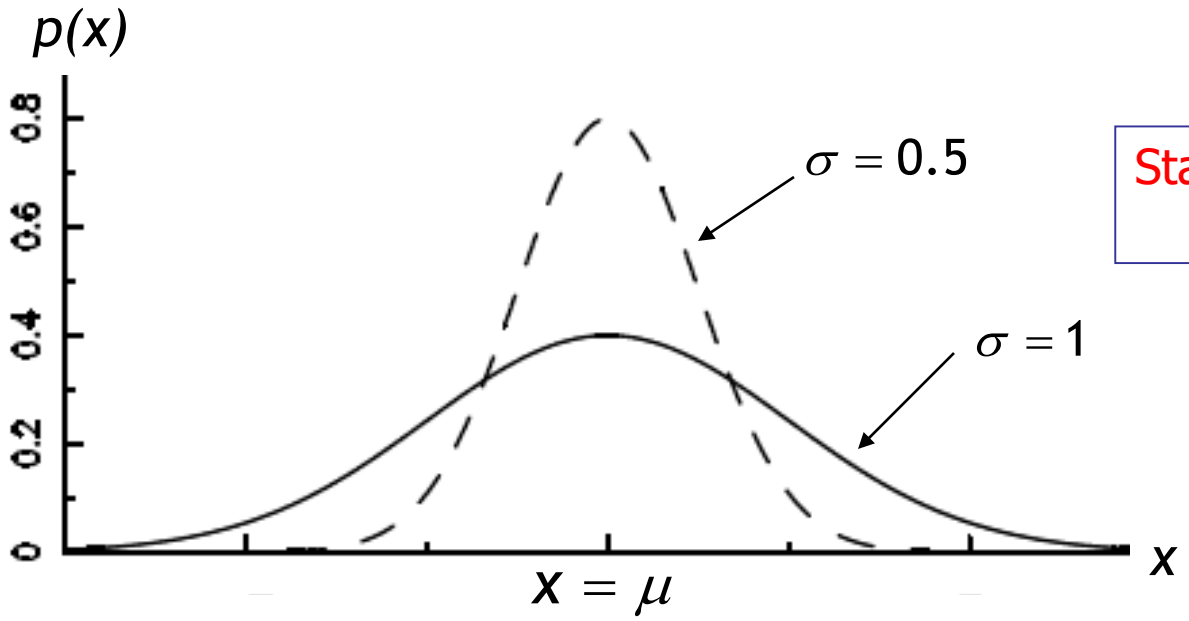
We get N photons for a source, the expected photons from background is M, then the source is detected at the significance of $(N-M)/N^{1/2}$ sigma level



- Mean: μ Variance: μ
- Approximate Gaussian distribution, as μ increases

Gaussian (normal) distribution

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



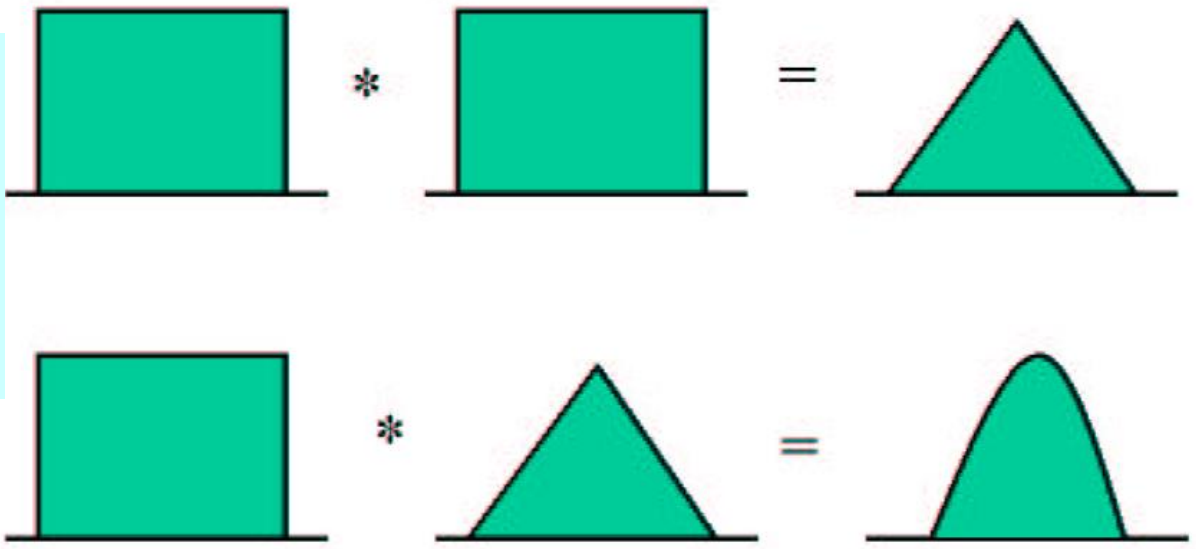
Standard normal
$\mu=0$ $\sigma=1$

Mean: μ Variance: σ^2

Why Gaussian: central limit theorem

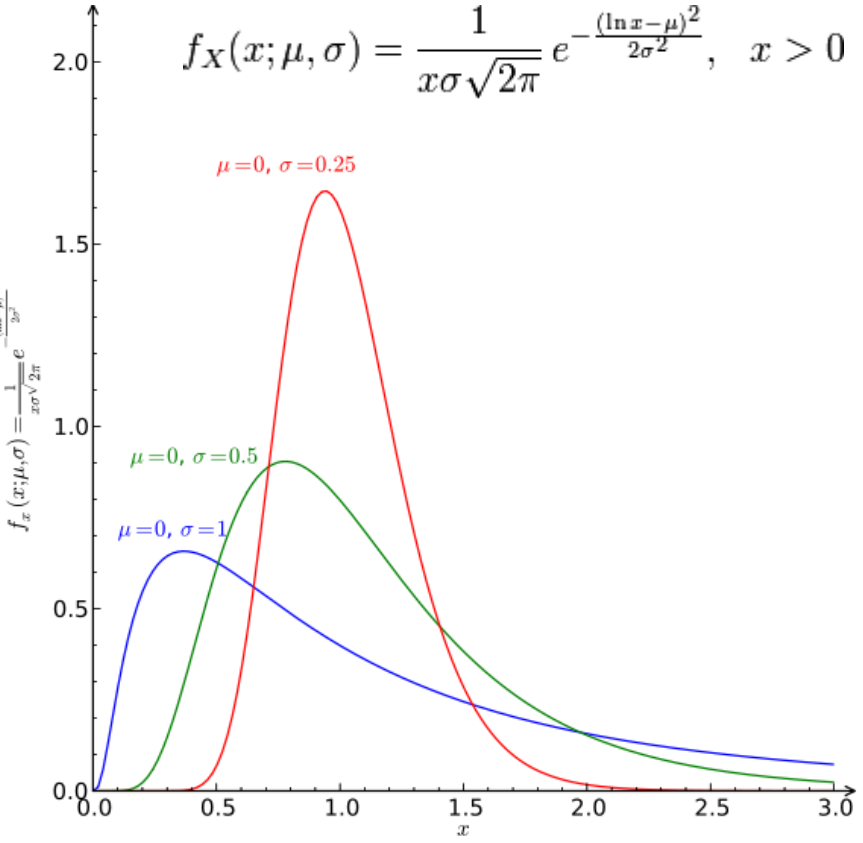
- The sum of n random values drawn from any probability distribution function of finite variance, σ^2, tends to be Gaussian distributed about the expectation value for the sum with variance $n\sigma^2$

The add of two independent random variables results a distributions that is the convolution of the two distribution functions



Why log-normal distribution?

- f(x)=G[ln(x)]
  - e.g. Concentration, spin of dark matter halo, galaxy size distribution etc.
- $f(x)=f_1*f_2*\ldots f_n$
  - $\log[f(x)]=\log f_1+\log f_2+\ldots \log f_n$
    - Central limit theory: Normal distribution



Conclusions of Central limit theorem

- The sampling distribution is known even when the underlying PDF is not
  - Sampling is a random process
  - The mean of a large sample tends to be normally distributed

- Under certain conditions, e.g. with so many unknown variables, we can assume an unknown distribution is Gaussian.
  - Distribution of human heights

More on Gaussian

- $G_1 + G_2$: Gaussian
- G1*G2 → log-normal
- $\mu 1/\mu_2$: Lorentz distribution
  - Infinite variance $p(x) = \frac{1}{\pi(1+x^2)}$
  - Appears in spectral line fitting
- $\mu_1^2 + \mu_2^2$: χ^2 (v=2)distribution

χ^2 distribution

χ^2 distribution with freedom of K is the a sum of the squares of K independent standard normal random variables.

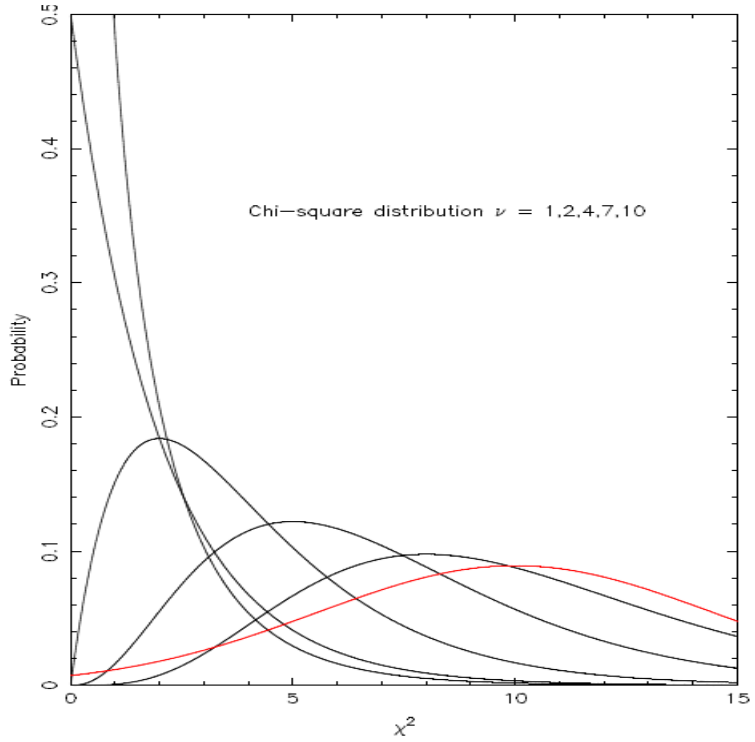


Figure 1: Examples of χ^2 distribution - N(10,20) Gaussian in red

$$P(x)= \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

Approximate Gaussian again when K is large

Mean: *K* Variance: *2K*

Student's t distribution t(ν)

Standard normal/χ^2 (ν)distribution

- Broader than Gaussian

- Used in check whether two distributions have the same mean:

$$t = \frac{\overline{x_A} - \overline{x_B}}{[\mathrm{Var}(x_A)/N_A + \mathrm{Var}(x_B)/N_B]^{1/2}}$$

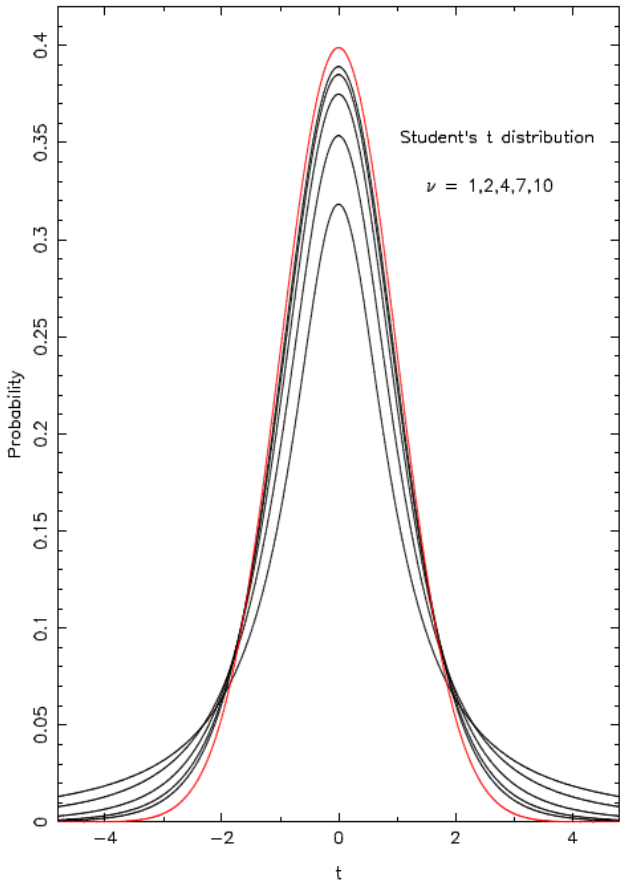
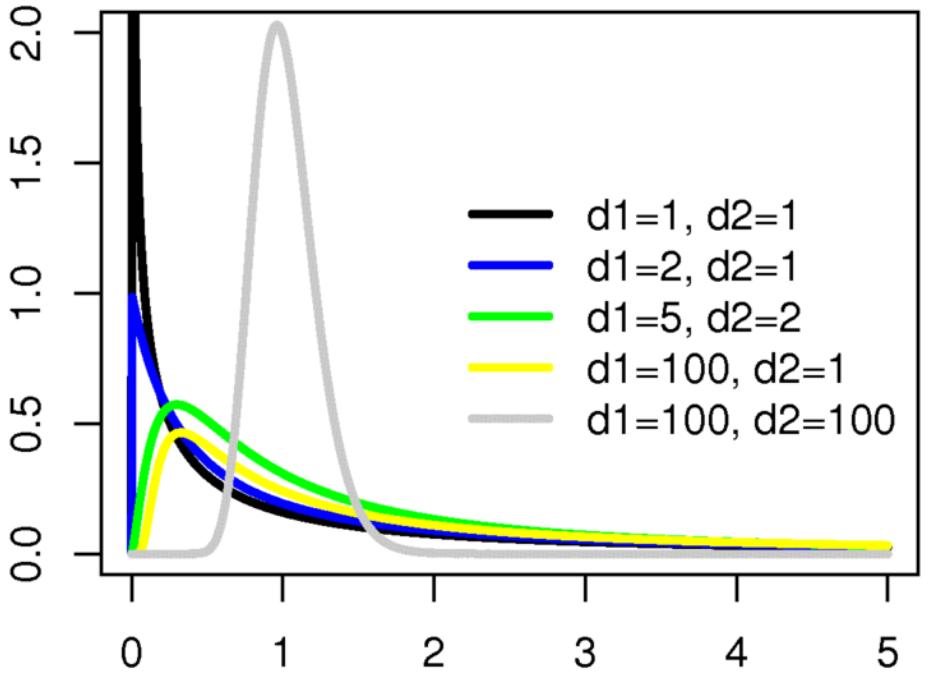
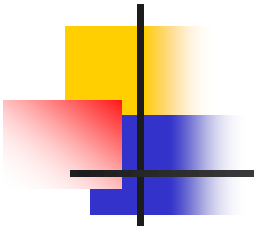


Figure 2: Examples of Student's t distribution - N(0,1) Gaussian in red

F - distribution

- Ratio of two χ^2 distributions F(d1,d2)
  - Check whether two distributions have the same variance





Modeling of data

- Is a hypothesis/model acceptable for given data
  - χ^2, t, F-tests
- For multiple models, which one describes data better?
  - Bayesian theorem
- For an assumed model with free parameters, what are the best estimation of the model parameters?
  - Minimum χ^2 fitting
  - Maximum likelihood estimation
- Non-parametric statistics
  - Are two samples drawn from the same distribution? (we don't know the distribution shape)
  - Do two parameters correlate? (we don't know how they correlate)

least square fits

Suppose that we are fitting N data points (x_i, y_i) $i = 1, \ldots, N$, to a model that has M adjustable parameters a_j, $j = 1, \ldots, M$. The model predicts a functional relationship between the measured independent and dependent variables,

$$y(x) = y(x; a_1 \ldots a_M) \tag{15.1.1}$$

where the dependence on the parameters is indicated explicitly on the right-hand side.

What, exactly, do we want to minimize to get fitted values for the a_j's? The first thing that comes to mind is the familiar least-squares fit,

$$\text{minimize over } a_1 \ldots a_M : \quad \sum_{i=1}^{N} [y_i - y(x_i; a_1 \ldots a_M)]^2 \tag{15.1.2}$$

least-squares fitting *is* a maximum likelihood estimation of the fitted parameters *if* the measurement errors are independent and normally distributed with constant standard deviation

Chi-square fitting

- Data points can not have the same error

$$\chi^2 \equiv \sum_{i=1}^{N} \left(\frac{y_i - y(x_i; a_1 \ldots a_M)}{\sigma_i} \right)^2$$

Understand as a weight of each data in least square fit

- Chi-square distribution with ν=N – M degrees of freedom
  - Model is perfect
  - Measurement error is right and Gaussian
  - Sample is not biased
- What does it mean if we get χ^2=30 for freedom 10?
  - P(χ^2 >30|10)=0.001: the probability we reject one of above assumptions is wrong (At least one of the above assumption is wrong at 99.9% level.)

Reasonable model

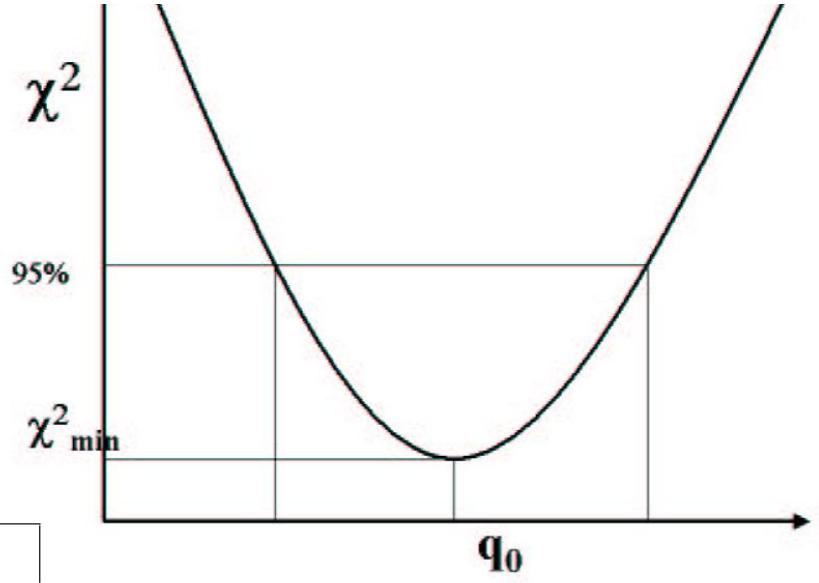
- If $\chi^2 \sim \nu$, model is reasonable
- If $\chi^2 << \nu$
  - Error is overestimated
  - Error is too large to distinguish models
- If $\chi^2 >> \nu$ (most of cases)
  - Ideal model never exists, e.g. Scaling relations
  - Data errors are underestimated
  - Data are always biased

freedom is a question! see arXiv:1012.3754

Minimum χ^2 and confidence level

For parameter sets, the one with minimum χ^2 is the best model, but may not be the correct one.

Compared to the best model, do other models also acceptable? What is the confidence level of the best estimation?

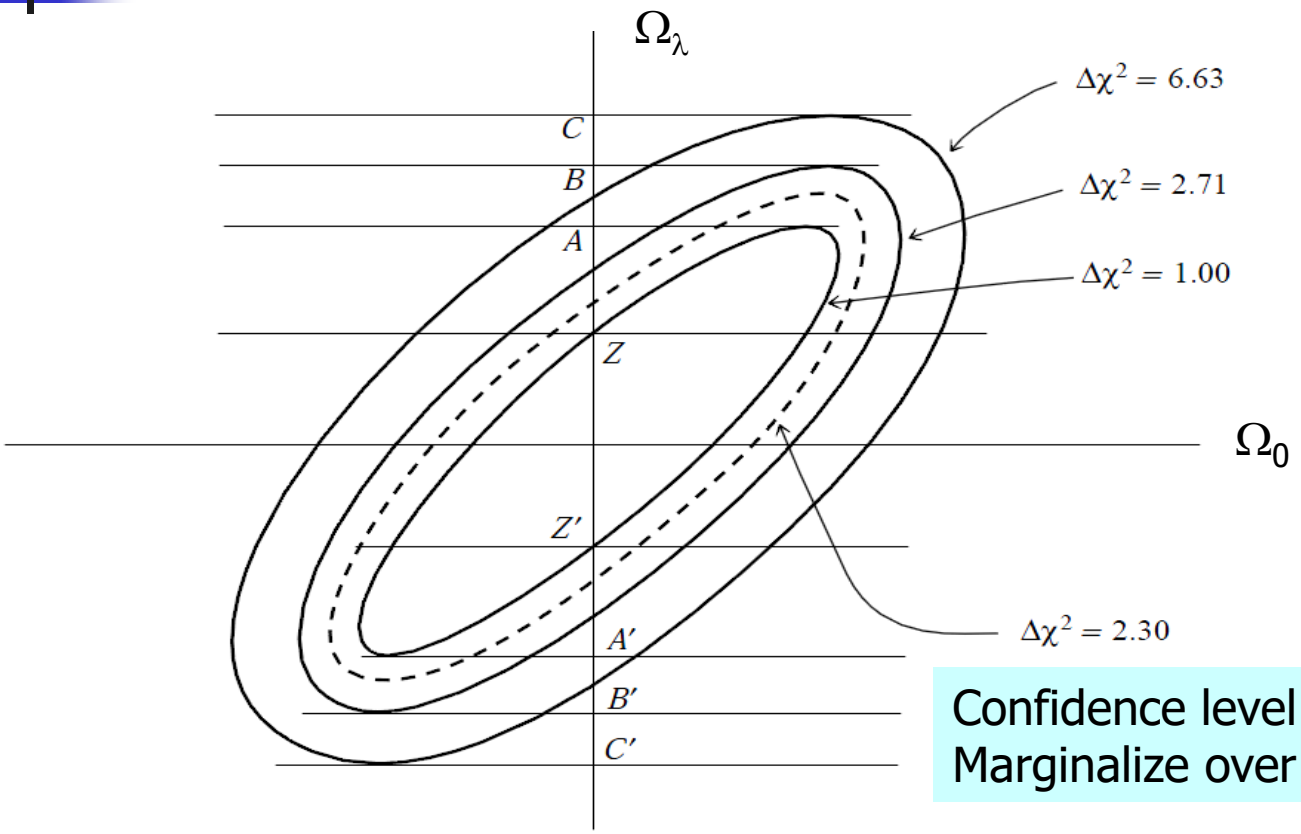


1: *Confidence regions for estimating Ω_V from supernova data.*

$\Delta\chi^2$ as a Function of Confidence Level and Degrees of Freedom						
	ν					
p	1	2	3	4	5	6
68.3%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.4%	4.00	6.17	8.02	9.70	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.8

- $\chi^2(q_0)=\chi^2_{min}+\Delta\chi^2$
- $\Delta\chi^2$ follows χ^2 distribution with M freedoms

Projections of the M-dimensional confidence regions



Confidence level of one parameter: Marginalize over other parameters

$$p(a|DI) = \int_0^\infty d\sigma p(a, \sigma|DI).$$

χ^2 test on binned data

- Compare the model prediction with the observed data in bins, N_i is the number of events observed, n_i is the number expected.
  - $P(N_i|n_i)$ is Poisson distribution, but approaches Gaussian when $n_i > 5$

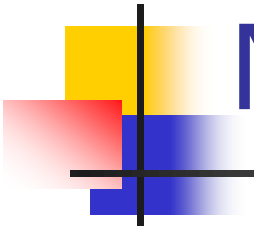
$$\chi^2 = \sum_i \frac{(N_i - n_i)^2}{n_i}$$

- Compare two data sets

Next we consider the case of comparing *two* binned data sets. Let R_i be the number of events in bin i for the first data set, S_i the number of events in the same bin i for the second data set. Then the chi-square statistic is

$$\chi^2 = \sum_i \frac{(R_i - S_i)^2}{R_i + S_i} \qquad (14.3.2)$$

Notes on binned data



- Binned data is a simple way to show the statistical properties of a sample
  - Why we always plot the histograms of the sample properties first.
- However, bins lose information
  - If we have some priors on the distribution function, we may do better.

Example on binned data

- We measured radial velocity of a emission line, and need to fit the velocity dispersion.

- We have ten data points (mean subtracted): 2.78, -1.84, 1.80, 0.11,-0.92, -0.91, 0.29, 3.57, -1.77, 2.55
  - M1: bin the measurement, plot the histogram, fit with a Gaussian profile
    - Not enough bins
  - M2: $P(v_i)=A \exp(-v_i^2/2\sigma^2)$
    - $\ln L = \Sigma \ln[P(v_i)] + \text{const} = -10\ln\sigma - (1/2\sigma^2) \Sigma v_i^2$
    - $d \ln L / d\sigma = 0 \rightarrow \sigma_{best}=1.97$
    - $\Delta \ln L=0.5$ (68% confidence level) σ-[1.60, 2.50]

Maximum likelihood (ML) estimation

- If the measured data are independent likelihood function

$$L(D_1,\cdots,D_n|\theta_1,\cdots,\theta_m) = p(D_1|\theta_1,\cdots,\theta_m)\cdots p(D_n|\theta_1,\cdots,\theta_m) = \prod_{i=1}^{n} p(D_i|\theta_1,\cdots,\theta_m).$$

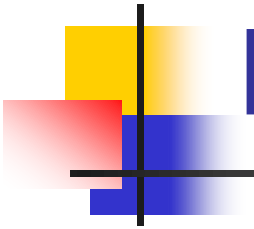
$$\frac{\partial}{\partial\theta}\ln L = 0.$$

- When the possibility distributions are Gaussian, the ML estimation is equivalent to minmum-χ2
  - Use $\Delta\chi^2$ to estimate the goodness of fit ($\Delta\ln L=-\Delta\chi^2/2$)
- For non-Gaussian likelihood function
  - we may quote where L is some fraction (e.g. 0.1) of L_{max}
- Fisher Information Matrix

$$F_{ij} = -\frac{\partial^2 \ln L}{\partial\theta_i\partial\theta_j}$$

- Monte-Carlo resampling

Notes on ML estimation



- ML is more general than minimum χ^2
  - e.g. we can calculate likelihood for detections with only upper limit
- Find out the ML is difficult and time-consuming when number of parameters is large
  - Monte-Carlo Markov chain(MCMC)

Bayesian Theorem

P(B|A)=P(B)*P(A/B)/P(A)

A: women B: pregnant

prior probability

Posteriori probability

$$p(H_i|DI) = \frac{p(H_i|I)p(D|H_iI)}{p(D|I)}.$$

evidence

Bayesian approach

- D: data, I: model, H: hypothesis (model parameter)

$$p(H_i|DI) = \frac{p(H_i|I)p(D|H_iI)}{p(D|I)}.$$

- Compare two hypotheses: H1 VS H2

$$\frac{p(H_1|DI)}{p(H_2|DI)} = \frac{p(H_1|I)p(D|H_1I)}{p(H_2|I)p(D|H_2I)}$$

  - in the absence of information, we could assume equal priors
  - p(D|H,I) is easy to get
  - Prior may change the possibility
    - e.g. P(t>13.6Gy)=0 for stellar population

Example: flux of GRB

Take Gamma Ray Bursts to be equally luminous events, distributed homogeneously in the Universe. We see three gamma ray photons from a GRB in an interval of 1 s. What is the flux of the source, *F*?

F=3 photons/s, with an uncertainty of about 1.73

Prior: low flux sources are intrinsically more probable, as there is more space for them to sit in. (Malmquist bias)

$$F = \frac{L}{4\pi r^2} \qquad \frac{dF}{dr} \propto -r^{-3} \qquad p(r \mid I)dr \propto 4\pi r^2 dr$$

$$p(F \mid I) \propto p(r \mid I)\left|\frac{dr}{dF}\right| \propto F^{-5/2}$$

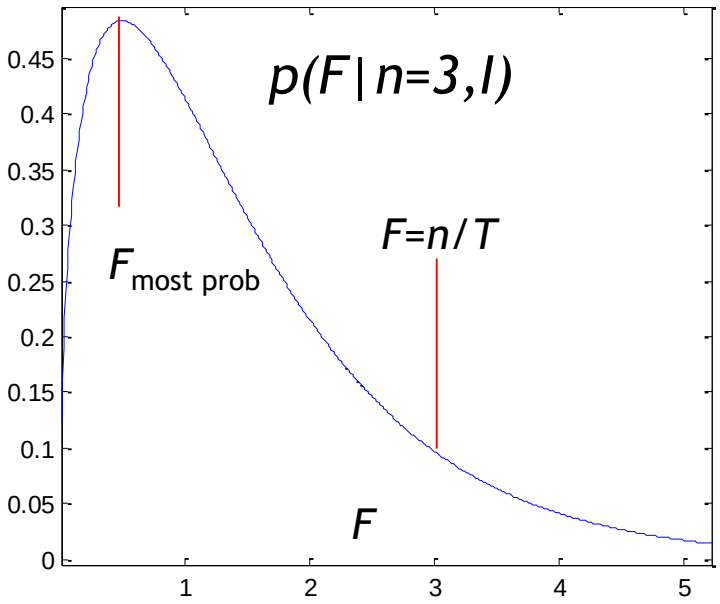
Bayesian estimation

- The posterior for *F* after seeing *n* photons

$$p(n \mid F, I) = F^n \exp(-F) / n!$$

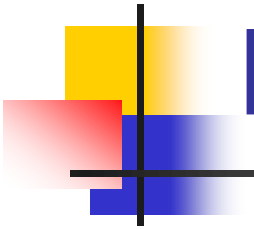
$$p(F \mid n, I) \propto F^{n-5/2} \exp(-F)$$

$$p(F \mid n, I) \propto p(F \mid I)\, p(n \mid F, I)$$



It is more probable this is a distant source from which we have seen an unusually high number of photons than it is an unusually nearby source from which we have seen an expected number of photons.

we get the most probable value of *F* equalling 0.5 photons/sec.



Non-parametric statistics

- The Kolmogorov-Smirnov test (K-S) test
  - Are two distributions different?
    - Even we do not know what is the distribution
- The Spearman rank correlation coefficient
  - Are two quantities correlated?
    - Not necessary linear

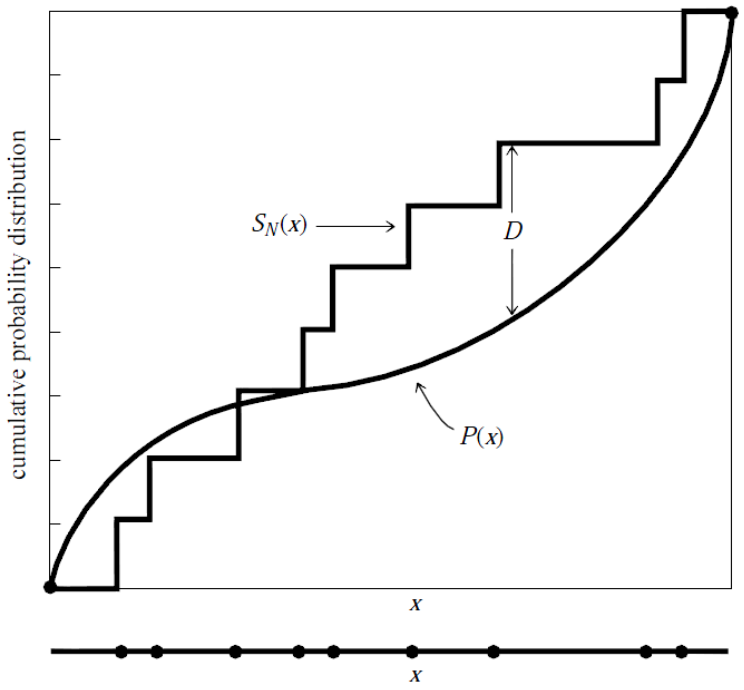
K-S test: applicable to unbinned distributions

- *K–S* test defined as the *maximum value* of the absolute difference between two cumulative distribution functions.

$$D = \max_{-\infty<x<\infty} |S_{N_1}(x) - S_{N_2}(x)|$$

$$P(> D) = 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2ni^2D^2}$$

- Can be generalized to two-dimensional distributions



- invariant of the parameterization of x
- most sensitive around the median

Spearman rank correlation coefficient

- Standard parametric (linear) correlation coefficient
  - -1<r<1

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

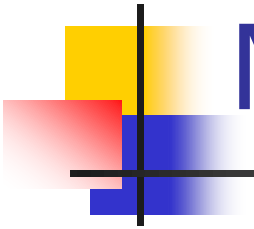
- Spearman rank: replacing x_i, y_i by the rank R_i, S_i

$$r_s = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2} \sqrt{\sum_i (S_i - \bar{S})^2}} \qquad (14.6.1)$$

The significance of a nonzero value of r_s is tested by computing

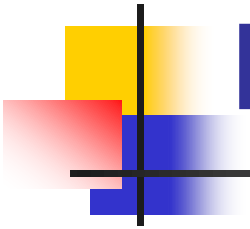
$$t = r_s \sqrt{\frac{N-2}{1-r_s^2}} \qquad (14.6.2)$$

t: Student's distribution with $N - 2$ degrees of freedom.



Monte-Carlo Methods

- Random sampling of distributions
  - Random number generator
- Error estimation: simulate the random process
  - Boot-strap
  - Jack-nife



Random generator

The very basic : random number
U[0,1]

Recipes for popular distributions
Gaussian, Poisson

Probability integral transform

Suppose we can compute the CDF of some desired random variable

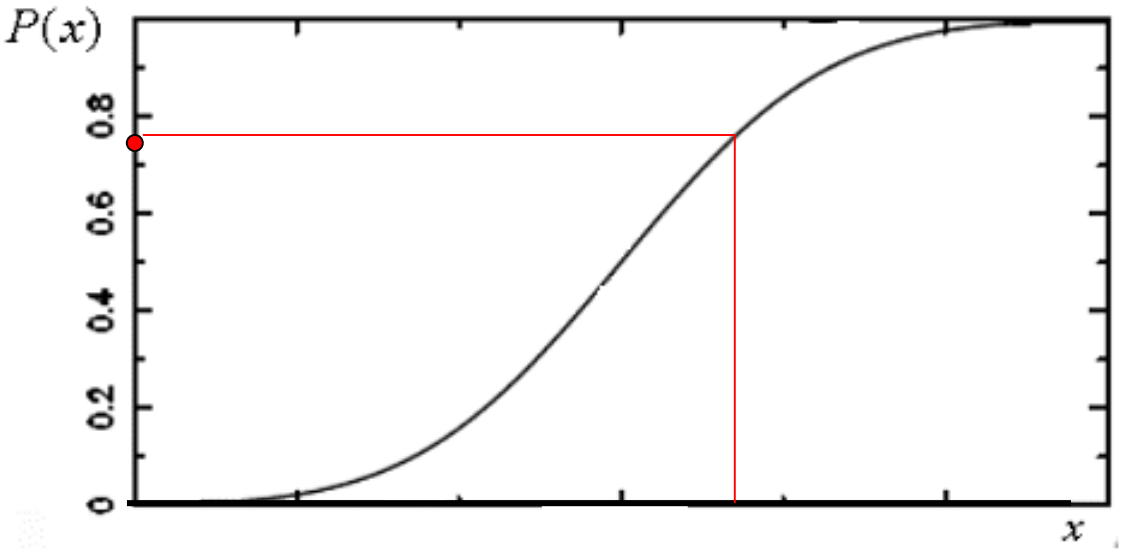
1) $y \sim U[0,1]$

2) Compute $x = P^{-1}(y)$

3) Then $x \sim p(x)$

Cumulative distribution function (CDF)

$$P(a) = \int_{-\infty}^{a} p(x)dx = \text{Prob}(x < a)$$



Rejection Sampling

Suppose we want to sample from some pdf $p(x)$ and we know that

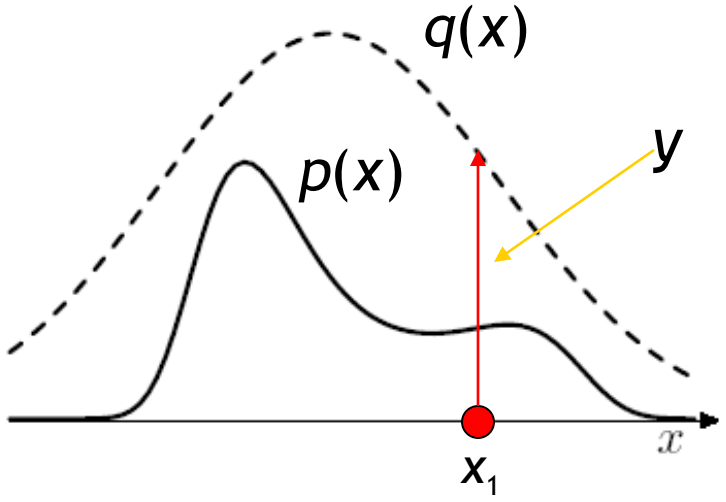
$$p(x) < q(x) \quad \forall x$$

where $q(x)$ is a simpler pdf called the *proposal distribution*

1) Sample x_1 from $q(x)$

2) Sample $y \sim U[0, q(x_1)]$

3) If $y < p(x)$ **ACCEPT**
   otherwise **REJECT**



Set of accepted values $\{x_i\}$ are a sample from $p(x)$.

Estimating error from data

- Error (confidence level) of the statistical parameter after complicated process
- Principle: probability is the frequency of the data
  - Sub-sample: N sub-samples

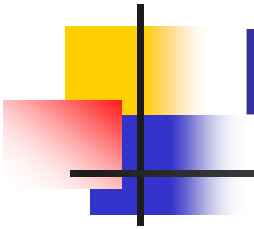
$$\sigma_{ii}^2 = \frac{1}{N-1}\sum_{k=1}^{N}\frac{({}^kP_i - \bar{P}_i)^2}{N},$$

  - Jack-Knife: omit each sub-sample in turn

$$\sigma_{ii}^2 = \frac{N-1}{N}\sum_{k=1}^{N}({}^kP_i - \bar{P}_i)^2,$$

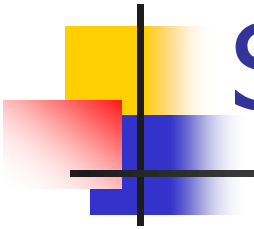
  - Boot-strap: resampling the same size sample

$$\sigma_{ii}^2 = \sum_{k=1}^{N}\frac{({}^kP_i - \bar{P}_i)^2}{N},$$



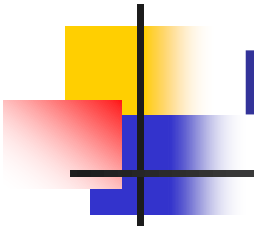
Notes

- Sub-sample method is problematic when the statistics needs the full sample
- Critical assumption for these MC methods is that the individual subsamples are independent
  - Not the case in the galaxy correlation function
    - Sub-samples large enough than the correlation length



Summary

- the basic probability distribution functions (moments)
  - e.g Gaussian, Poisson
- the basics of the data
  - Are the data biased?
  - How about the error?
    - Error propagation
    - Numerical error estimation (e.g. boot-strap)



more

- the basics of the modeling data
  - Distribution of one quantity
    - Histograms (always first step)
    - Guess a function and parameterize it
    - Maximum likelihood estimation of the parameters
  - Scaling relations between quantities
    - linear correlation (always first step)
  - Build model: motivated by physics
    - Estimate model parameters from data