STATISTICS IN ASTRONOMY

- How can data be used best/optimally ?
- Error assignment ? What does it mean ?
  Confidence intervals ?
- Model fitting = testing hypotheses = what parameters ?
- Correlation between variables ? Are objects distributed randomly ? Clustered ?
- Limits on knowledge = imperfect equipment, sample size
- Underlying distributions frequently non-Normal and often **unknown**
- How to deal with noise outliers and unknown errors?
- Non-parametric methods; why do we need them ?
- How to group objects in classes ? Classify new objects ?

A FEW SIMPLE? PROBLEMS

- Everyone's dilemma

  How do you make best use of prior knowledge, or information ?

- The gambler's dilemma

  You observe **a** successes and **b** failures in **a** + **b** trials, what is the probability of **c** successes and **d** failures in **c** + **d** further trials ?

- The astronomer's dilemma I

  Example - the first Hubble Diagram = errors on both V and R

  How to calculate Ho ?

- The astronomer's dilemma II

  Example - Satellites of the Milky Way = small sample size;

  can you compute the form of the radial distribution ?

  To bin or not to bin ?

BAYES' THEOREM AND PRIOR KNOWLEDGE

You are unlucky enough to:-

- have had your DNA typed and recorded in a databank of ten million punters
- your DNA type matched 'perfectly' with a sample taken at the scene of a serious crime

A government scientist said "this test is virtually infallible with less than a one in million chance of giving the wrong answer"

Are you worried ?

$$P(false\ alarm) = P(DNA|innoc) = 10^{-6}$$

$$P(guilt|DNA) = \frac{P(DNA|guilt)P(guilt)}{P(DNA|guilt)P(guilt) + P(DNA|innoc)P(innoc)}$$

LEAST SQUARES – REGRESSION

Hubble law, $v = H_o\ d$, as example of $y = a\ x$ problem

$$y_i = a\ x_i + \epsilon_i$$

where y_i dependent & x_i independent variables; ϵ_i is error

$$\hat{a} = \sum_i x_i\ y_i / \sum_i x_i^2$$

What is error in $\hat{a}$?

$$var(\hat{a}) = \sigma_a^2 = < \hat{a}^2 > - < \hat{a} >^2$$

$$= < \epsilon_i^2 > / \sum_i x_i^2 \ = < \epsilon_i^2 > /N < x_i^2 >$$

But now suppose error in x_i as well !

$$x_i = \hat{x}_i + \epsilon_{x_i} \qquad y_i = a\ \hat{x}_i + \epsilon_{y_i}$$

$$\hat{a} = \sum_i x_i\ y_i / \sum_i x_i^2 \qquad < x_i^2 >_i = < \hat{x}_i^2 >_i + < \epsilon_{x_i}^2 >_i$$

$$\frac{1}{\hat{a}} = \sum_i x_i\ y_i / \sum_i y_i^2 \qquad < y_i^2 >_i = < \hat{y}_i^2 >_i + < \epsilon_{y_i}^2 >_i$$

PROBABILITY FUNCTIONS–I

Probability density function (PDF) $P(x)$, then $P(x)dx$ probability of x being in the range $x \rightarrow x + dx$

$$P(x) \geq 0 \;\; ; \;\; \int P(x)dx = 1$$

Cumulative probability function $C(x)$ probability of x being $\leq x$

$$C(x) = \int_{-\infty}^{x} P(y)dy$$

Characteristic function is the Fourier transform of the PDF

$$\phi(t) = \int e^{ixt} P(x)dx = < e^{ixt} >$$

Expanding the above equation

$$\phi(t) = < 1 + x(it) + \frac{1}{2!} x(it)^2 + ... > = 1 + \sum_r \frac{1}{r!} \mu_r' (it)^r$$

Which is none other than the moment generating function since

$$\frac{\partial^n \phi(t)}{\partial^n t}|_{t=0} \;\; = i^n < x^n > = i^n \mu_n'$$

PROBABILITY FUNCTIONS–II

Variable change, $y = f(x)$, probability invariant over interval, hence

$$P(y)\ dy = P(x)\ dx\ ; \Rightarrow \quad P(y) = P(x)\ |\frac{dy}{dx}|^{-1}$$

Joint distribution of two or more variables

$$P(x, y)\ dxdy \quad if\ independent \quad = \ f(x)dx\ g(y)dy$$

related to conditional distribution by

$$P(x, y)\ dxdy = P(x|y)\ P(y)\ dxdy$$

and the marginal distribution which is defined as

$$\phi(x) = \int P(x, y)\ dy$$

BINOMIAL AND MULTINOMIAL DISTRIBUTIONS

Binomial distribution gives probability of r successes in n trials

$$P(r) = ^n C_r \; p^r \; q^{n-r} = \frac{n!}{r!(n-r)!} \; p^r \; q^{n-r}$$

Rewrite this in terms of two outcomes n_1, n_2 with probabilities p_1, p_2

$$P(n_1, n_2) = \frac{n!}{n_1! \; n_2!} \;\; p_1^{n_1} \; p_2^{n_2}$$

Now generalise to m outcomes $\Rightarrow$ the Multinomial distribution

$$P(n_1, n_2, ...n_m) = \frac{n!}{n_1! \; n_2!......n_m!} \;\; p_1^{n_1} \; p_2^{n_2}........p_m^{n_m}$$

$$= \frac{n!}{\prod_{i=1}^{m} n_i!} \;\; \prod_{i=1}^{m} p_i^{n_i}$$

POISSON DISTRIBUTION

Random independent events eg. photon arrival at a detector

$$P(t)dt = \lambda dt$$

Probability of n events in time $0 \rightarrow t$ is given by

$$P(n,t) = \frac{(\lambda t)^n exp(-\lambda t)}{n!}$$

Poisson distribution for which mean and variance $\mu = \lambda t$; $\sigma^2 = \lambda t$

Characteristic function ϕ is given by

$$\phi = e^{\mu(e^{it}-1)}$$

Reminder: central moments μ_i are defined using

$$\mu_i = \int (x-\mu)^i P(x)dx$$

Example: CCD – if mean count in image is $N = \lambda t$

$$\sigma = \sqrt{N}$$

More generally, in pixel i, if signal s_i counts/s, background b_i counts/s and integrate for time t, what happens to the signal:to:noise, $s : n$, if the detector readout noise is r ?

$$s : n = \frac{s_i \times t}{\sqrt{(s_i + b_i) \times t + r^2}}$$

Two special cases:-

$$\text{Detector noise limited} \rightarrow \frac{s_i \times t}{r}$$

$$\text{Sky noise limited} \rightarrow \frac{s_i}{\sqrt{s_i + b_i}} \times \sqrt{t}$$

Now suppose the total flux/s in an image is given by $F = \sum_{i=1}^{N_{pix}} s_i$ then the equivalent expression for the total signal:noise is

$$s : n = \frac{F \times t}{\sqrt{F \times t + N_{pix}(b \times t + r^2)}}$$

Now three special cases:-

$$\text{Bright objects} \rightarrow \sqrt{F \times t}$$

$$\text{Sky limited} \rightarrow F\sqrt{t} \;/ \sqrt{N_{pix} \times b}$$

$$\text{Read noise} \rightarrow F \times t \;/\; \sqrt{N_{pix}\; r^2}$$

GAUSSIAN DISTRIBUTION – NORMAL – N(μ, σ^2)

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

This has a characteristic function ϕ given by

$$\phi = e^{it\mu - t^2\sigma^2/2}$$

Dominant in experimentation, probability theory and statistics because:

1. Approximated from Poisson for "large" λt

$$\mu = \sigma^2 = \lambda t \gtrsim 10$$

2. Central Limit Theorem
3. Distribution with maximum randomness (entropy) for given σ^2
4. Analytically tractable

STATISTICS

Statistics are functions of the data alone, for example any

$$s = f(x_1, x_2, ...x_i.....x_N)$$

is a statistic.

Useful statistics are **consistent** ie. they converge to the expectation value as the sample size increases and (usually) **unbiased** ie. on *average* they give the correct answer.

The following examples are consistent unbiased estimators of mean and variance

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \hat{\mu})^2$$

The variance on the estimate of the mean and the variance/ standard deviation is for uncorrelated variables

$$var(\hat{\mu}) = \frac{1}{N}\hat{\sigma}^2$$

$$var(\hat{\sigma}^2) \approx \frac{2(N-1)}{N^2}\hat{\sigma}^4$$

$$var(\hat{\sigma}) \approx \frac{1}{2N}\hat{\sigma}^2$$

SAMPLING – ESTIMATING CENTRAL LOCATION

Consider a series of measurements of some quantity, x_i, where $i = 1, m$. For any PDF, $P(x)$, various estimators of "central location" can be defined:

- MEAN minimises

$$< (x_i - \hat{x})^2 >_i = \int (x - \hat{x})^2 P(x).dx$$

- MEDIAN minimises

$$< |x_i - \hat{x}| >_i = \int |x - \hat{x}| P(x).dx$$

- MODE estimates position of maximum of $P(x)$

- Iff $P(x)$ exactly Gaussian (or Poisson) with variance σ^2, MEAN is optimum estimator with error $\sigma/\sqrt{m}$.
  MEDIAN has error $\sqrt{\pi/2} \times \sigma/\sqrt{m}$.

- Is there a Maximum Likelihood Estimator for a "real" PDF ?

LOCATION AND SCALE ESTIMATION

Gaussian distribution

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

Cauchy distribution

$$P(x) = \frac{1}{\pi\sigma}\ \frac{1}{1+(x-\mu)^2/\sigma^2}$$

Both are examples of a generic form for unimodal distributions

$$P(x)dx = f(\frac{x-\mu}{\sigma})\ \frac{dx}{\sigma}$$

where μ is a location parameter and σ is a scale parameter

Methods for estimation of scale (scatter) (spread) (sigma)

- $rms = < (x-\hat{\mu})^2 >^{1/2}$
- modulus $= < |x-\hat{\mu}| >$
- FWHM = full width at half-maximum
- interquartile range $\int_{x_l}^{x_h} P(x)dx = 1/2$
- MAD = median of the absolute deviation from the median

χ^2 DISTRIBUTION

Distribution of sum of squares of independent N(0,1) variates

$$\chi^2 = \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} \frac{(d_i - m_i)^2}{\sigma_i^2}$$

$$P(\chi^2) = \frac{1}{2^{\nu/2}(\nu/2 - 1)!} e^{-\chi^2/2} \chi^{\nu-2}$$

where $\nu =$ the number of degrees of freedom

$$< \chi^2 >= \mu = \nu$$

$$var(\chi^2) = \sigma^2 = 2\nu$$

In the limit as $\nu \to \infty$

$$P(\chi^2) \to N(\mu, \sigma^2)$$

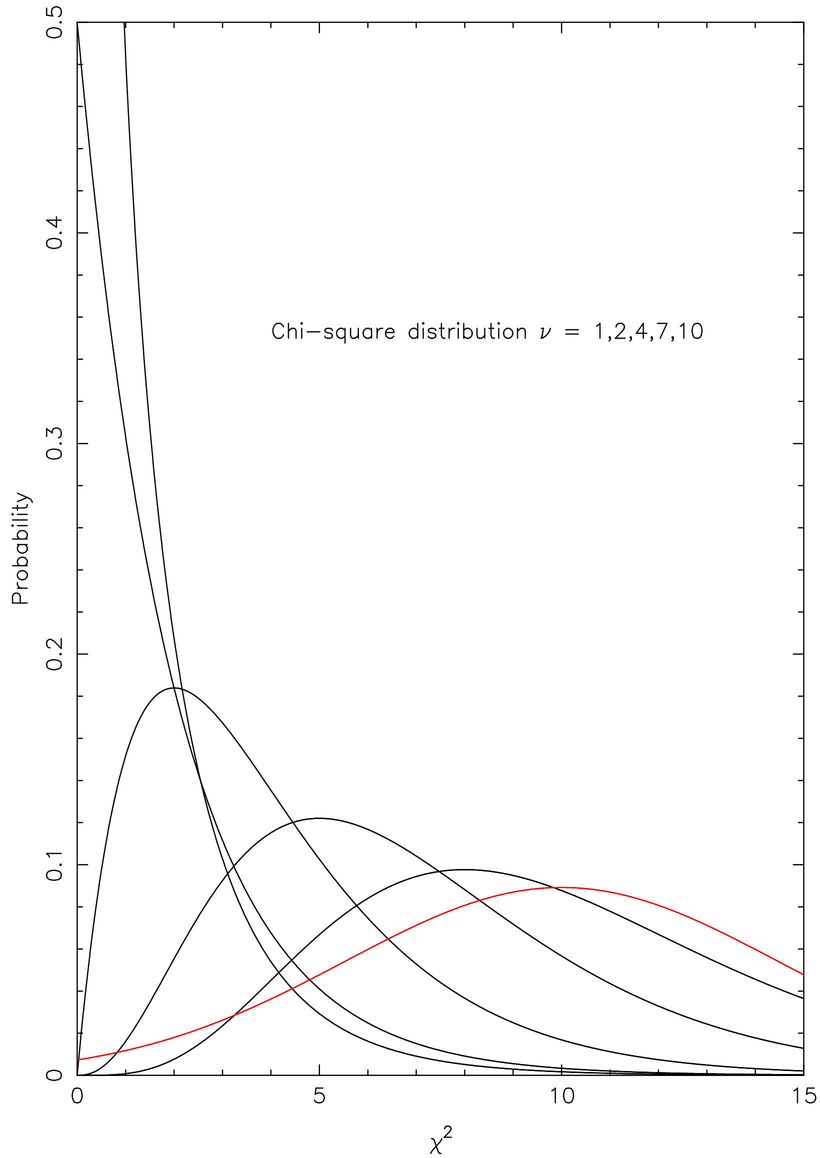

Figure 1: Examples of χ^2 distribution - N(10,20) Gaussian in red

STUDENT'S – t DISTRIBUTION

How many standard deviations is an estimate of the mean from the true value, μ, if both the mean $\hat{x}$ and the standard deviation $\hat{\sigma}$ are estimated from the data sample ?

Start by defining a suitably normalised variable

$$t = \frac{\hat{x} - \mu}{\hat{\sigma}/\sqrt{N}}$$

Then its PDF is given by

$$P(t) = \frac{(\nu/2 - 1/2)!}{\sqrt{\pi\nu}(\nu/2 - 1)!} \frac{1}{(1 + t^2/\nu)^{(\nu+1)/2}}$$

where $\nu = N - 1$ is the no. of degrees of freedom

$$< t > = 0 \qquad var\{t\} = \frac{\nu}{\nu - 2}$$

In the limit of large ν Student's - t $\rightarrow N(0,1)$

["Student" = W.S.Gosset = brewer for Guinness]

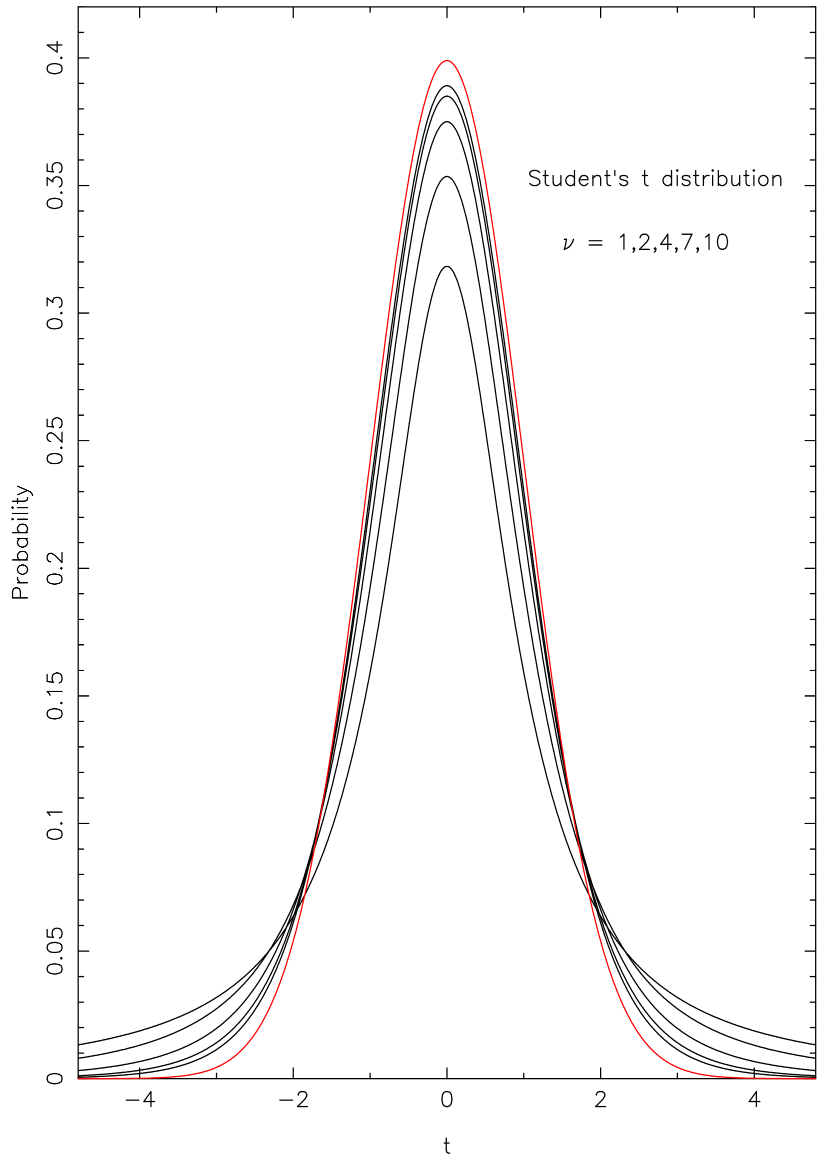

Figure 2: Examples of Student's t distribution - N(0,1) Gaussian in red